Statistica Sinica **31** (2021), 1749-1777
doi:https://doi.org/10.5705/ss.202018.0462

ON SURE SCREENING WITH MULTIPLE RESPONSES

Di He, Yong Zhou and Hui Zou

*Nanjing University, East China Normal University
and University of Minnesota*

Abstract: Multivariate responses are commonly encountered in many applications with high-dimensional input variables. Feature screening has been shown to be a very useful data analysis tool for high-dimensional data. Since the introduction of the sure independence screening approach, many variable screening methods have been proposed and studied in the literature. However, the majority of these methods focus on the classical univariate response data case, and do not apply naturally to data sets with multiple responses. We systematically study variable screening methods for multi-response data. First, we consider extensions of several popular screening methods to deal with multiple responses. Each of these methods has its own clear drawbacks. We then propose a new model-free screening method, which we call multi-response rank canonical correlation screening (mRCC), which not only takes into account the dependence structure among the multivariate responses, but also preserves nice properties of the rank correlation, such as robustness and invariance under monotonic transformation. The sure screening property of mRCC is established under weak regularity conditions. Extensive numerical experiments demonstrate the superior performance of mRCC over other available alternatives.

Key words and phrases: Canonical correlation, multi-response data, rank correlation, sure screening property.

1. Introduction

Multivariate responses are commonly encountered in many statistical applications. For example, microarray expression experiments and array comparative genomic hybridization (CGH) experiments have been conducted by biologists in breast cancer cohort studies (Sørlie et al. (2001); Zhao et al. (2004); Chin et al. (2006); Bergamaschi et al. (2008)). The resulting data from these experiments are RNA transcript levels and DNA copy numbers. Although analyses of expression arrays alone or CGH arrays alone have provided useful information, an integrative analysis of DNA copy numbers and gene expression files is necessary, because these two types of data offer complementary information. Hence, integrating DNA and RNA data benefits the recognition of more subtle genetic

Corresponding author: Hui Zou, School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA. E-mail: zouxx019@umn.edu.

regulatory relationships in cancer cells (Pollack et al. (2002)).

A straightforward way to model and analyze such data sets is to use a multi-response regression, though our method is not limited to the regression model. Let n denote the sample size, p the number of predictors, and q the number of responses. A multi-response regression model is

$$\mathbf{Y} = \mathbf{B}_0 + \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{1.1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_q)$ is an $n \times q$ response matrix, $\mathbf{X} = (X_1, \ldots, X_p)$ is an $n \times p$ design matrix, $\mathbf{B} = (\beta_{kj})$ is a $p \times q$ matrix of parameters, $\mathbf{B}_0 = (\beta_{01}\mathbf{1}, \ldots, \beta_{0q}\mathbf{1})$ is an $n \times q$ matrix of intercepts, with $\mathbf{1}$ an n-vector with all entries equal to one, and $\mathbf{E}$ is an unobserved $n \times q$ matrix, with row vectors $\epsilon_1, \ldots, \epsilon_n$ that are independent copies with mean zero and covariance matrix $\mathbf{\Sigma_E}$. In general, we should not treat a multi-response problem as multiple univariate response problems, although the solutions may sometimes be the same. For example, we can obtain the ordinary least squares estimator of (1.1) by performing a separate linear regression on each response. If the errors are correlated, a weighted criterion of the residual sum of squares arises naturally, and the solution still amounts to the ordinary least squares estimates. However, this is not the case for a regression with a Lasso penalty on the entries of $\mathbf{B}$. When a Lasso regression involves a known $\mathbf{\Sigma_E}$, the optimal solution for $\mathbf{B}$ obtained from the weighted criterion accounts for the inverse of $\mathbf{\Sigma_E}$ (Rothman, Levina and Zhu (2010)), which is different from the separate Lasso regression estimates with each response. When p is very large, there are challenges related to computational efficiency, statistical consistency, and algorithmic stability (Fan, Samworth and Wu (2009)). To this end, many shrinkage estimators of the parameters have been proposed for the multi-response regression in (1.1) that penalize the optimization with the residual sum of squares. Some simultaneously estimate the parameters and discard irrelevant predictors using proper regularization (Obozinski, Taskar and Jordan (2010); Peng et al. (2010); Lee and Liu (2012)). Others encourage an estimator of reduced rank (Anderson (1951); Yuan et al. (2007); Chen and Huang (2012)), in which dimension reduction is achieved by constraining the coefficient matrix to have low rank.

Fan and Lv (2008) argue that it is beneficial for both computations and theoretical considerations to first reduce the ambient dimension to a moderately high dimension, and then to fit a regularized model. The dimension-reduction step should preserve all important features–a property known as the sure screening property. To demonstrate their philosophy, Fan and Lv (2008) introduced a sure independence screening (SIS) procedure, using a Pearson correlation to filter out a large number of noise variables. SIS is shown to have the sure screening property. Inspired by this influential paper, many researchers have studied the variable screening problem and proposed more sophisticated screening methods to deal with more complicated models. These include maximum marginal likelihood screening for generalized linear models (Fan and Song (2010)), nonparametric independence screening (NIS) for additive models (Fan, Feng and Song (2011)), robust rank correlation screening (RRCS) for semiparametric single-index models with a monotonic link function (Li et al. (2012)), quantile-adaptive screening for a quantile regression (He, Wang and Hong (2013)), empirical likelihood screening for parametric models that can be formulated using general estimating equations (Chang, Tang and Wu (2013)), and so on. Fan, Ma and Dai (2014) extended NIS for varying-coefficient models, and Liu, Li and Wu (2014) considered these types of models based on a conditional correlation coefficient. Chang, Tang and Wu (2016) proposed a unified approach for nonparametric and semiparametric models based on the marginal empirical likelihood. When the response is binary, Fan and Fan (2008) proposed a t-statistic to screen predictors, and Mai and Zou (2013) developed the Kolmogorov filter using the Kolmogorov–Smirnov statistic. Huang, Li and Wang (2014) proposed a Pearson chi-square-based feature screening method for categorical responses and predictors. Cui, Li and Zhong (2015) considered a discriminant analysis with a multi-categorical response variable. Another popular screening genre is that of the model-free methods, which overcome the model misspecification problem. For instance, these methods include the sure independent ranking and screening (SIRS) (Zhu et al. (2011)), distance correlation screening (DCS) (Li, Zhong and Zhu (2012)), and fused Kolmogorov filter (Mai and Zou (2015)).

The focus of this study is multi-response data. The aforementioned screening methods are primarily developed for the univariate data case. To the best of our knowledge, the only method that can naturally handle multivariate and univariate responses is the distance correlation screening (DCS) method because a distance correlation can be defined between two random vectors. However, it has been observed that in the presence of heavy-tailed data, the performance of DCS can be very poor (Mai and Zou (2015)). This is because the sure screening property of DCS relies on a moment condition that the response and the predictors should be sub-Gaussian. When the assumption is violated, the sure screening property of DCS becomes questionable, limiting its application to the multivariate responses. Furthermore, the DCS is not invariant against monotonic transformation.

The main goal of this study is to develop new variable screening methods for multi-response data. First, we extend several existing screening methods (SIS, NIS, RRCS) to the multi-response case by simply summing up the squares of the marginal utility with every component of the multivariate response, which is equivalent to treating the problem as multiple univariate response data problems. We believe a better variable screening method is possible if we consider potential dependence between multiple responses. We propose a new approach called multi-response rank canonical correlation screening (mRCC) without imposing a model assumption. This new model-free method integrates two commonly used rank correlations, Spearman's correlation and/or Kendall's τ correlation, with a canonical correlation. This inherits the multivariate merits of the canonical correlation that takes advantage of the dependence structure among the multivariate responses. In addition, it preserves nice properties of the rank correlation that can handle heavy-tailed predictors and responses, as well as invariance against monotonic transformations. Moreover, mRCC is easy to implement and cheap to compute. The sure screening property can be shown under very weak conditions, without assuming any moment conditions on the predictors and responses. Hence, we recommend using the mRCC method for variable screening with multi-response data.

The rest of the paper is organized as follows. Extensions of several existing screening methods are given in Section 2. In Section 3, we first introduce a screening method based on a canonical correlation, and then propose mRCC. The theoretical discussion in Section 4 shows that the sure screening property of mRCC holds under weak regularity conditions. Section 5 presents our simulation experiments and a genomic data example, which we use to compare the methods. The technical proofs are presented in the Appendix.

Notation. Throughout this paper, we assume $\mathbf{X}$ is centered to have mean zero columnwise. We denote X_k, Y_j as the kth, jth column of $\mathbf{X}, \mathbf{Y}$, for $k = 1, \ldots, p$ and $j = 1, \ldots, q$. To avoid introducing additional notation, we sometimes refer to X_k, Y_j or X, Y as the vectors of the samples, and sometimes refer to them as the random variables, when necessary. We also abuse $\mathbf{Y}$ to denote the random vector of the response. Let $\|\cdot\|$ be the Euclidean norm for a vector, and let $\|\cdot\|_{\mathrm{F}}$ be the Frobenius norm for a matrix. Denote RSS as the residual sum of squares from a regression.

2. Extensions of Existing Screening Methods

2.1. Sure independence screening

Fan and Lv (2008) proposed the SIS method, using the marginal correlation ranking $X_k^{\mathrm{T}} Y / \|X_k\| \|Y\|$ to filter out features that are weakly correlated with the response. SIS can be viewed from a marginal regression perspective:

$$\min_{\beta_0, \beta_k} \left\| Y - \beta_0 \mathbf{1} - X_k \beta_k \right\|^2. \tag{2.1}$$

Under the condition that the X_k are further standardized to have norm one, it is easy to show that this is equivalent to ranking by the absolute value of the regression coefficient, by the magnitude of the Pearson correlation coefficient, or by the descending order of the RSS of the marginal regression. To carry out a similar screening procedure when the response is multivariate, a straightforward idea is to generalize (2.1) to

$$\min_{\mathbf{B}_0, \boldsymbol{\beta}_k} \left\| \mathbf{Y} - \mathbf{B}_0 - X_k \boldsymbol{\beta}_k \right\|_{\mathrm{F}}^2, \tag{2.2}$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kq})$ is a row vector of parameters. The RSS has the following form:

$$RSS_k = \sum_{j=1}^{q} \|Y_{j_c}\|^2 \cdot \left(1 - \widehat{\rho}_{kj}^2\right),$$

where $Y_{j_c} = Y_j - \bar{Y}_j \mathbf{1}$, and $\widehat{\rho}_{kj} = \widehat{\rho}(X_k, Y_j)$ is the sample Pearson correlation coefficient between X_k and Y_j. We scale $\mathbf{Y}$ and $\mathbf{X}$ to have mean zero and norm one columnwise in order to remove the scale influence. In this case, the rankings according to the following three quantities are still equivalent: the ℓ_2-norm of the coefficient vector, the sum of the squares of the Pearson correlation coefficients, and the descending order of the RSS. Hence, by aggregating the squares of the Pearson correlation coefficients of the predictor with each response, we obtain

$$\widehat{\omega}_k^{\mathrm{mSIS}} = \|\widehat{\boldsymbol{\rho}}_k\|^2, \tag{2.3}$$

where $\widehat{\boldsymbol{\rho}}_k = (\widehat{\rho}_{k1}, \ldots, \widehat{\rho}_{kq})^{\mathrm{T}}$, and refer to (2.3) as the multi-response sure independence screening (mSIS) statistic. Note that this approach actually treats the multi-response problem as multiple univariate response data problems. It has been observed that SIS can fail when the linear regression model assumption does not hold for the data. It is expected that mSIS inherits this serious drawback of SIS.

2.2. Nonparametric independence screening

Fan, Feng and Song (2011) developed nonparametric independence screening

(NIS) for additive models, which allows the true regression function to be nonlinear in the predictors. They considered a marginal nonparametric regression using a basis function expansion such as B-splines. Similarly to the generalization of SIS, we aggregate RSS from the marginal nonparametric regressions with each response

$$\min_{\mathbf{f}_k \in \mathcal{S}_n^q} \left\|\mathbf{Y} - \mathbf{f}_k(X_k)\right\|_{\mathrm{F}}^2 = \min_{\mathbf{b}_k \in \mathbb{R}^{d_n \times q}} \left\|\mathbf{Y} - \mathbf{\Psi}_k \mathbf{b}_k\right\|_{\mathrm{F}}^2, \tag{2.4}$$

where $\mathbf{f}_k = (f_{k1}, \ldots, f_{kq})$, with $f_{kj}(X_k) = \sum_{l=1}^{d_n} \gamma_{kjl} \Psi_l(X_k)$ an n-vector sample version intending to approximate $E(Y_j|X_k)$, $\mathcal{S}_n$ is the space of polynomial splines, $\mathbf{\Psi}_k \triangleq (\Psi_1(X_k), \ldots, \Psi_{d_n}(X_k))$ denotes an $n \times d_n$ normalized B-spline basis matrix, $\mathbf{b}_k = (\boldsymbol{\gamma}_{k1}, \ldots, \boldsymbol{\gamma}_{kq})$, and $\boldsymbol{\gamma}_{kj} = (\gamma_{kj1}, \ldots, \gamma_{kjd_n})^{\mathrm{T}}$, for $j = 1, \ldots, q$. The corresponding solution is

$$\widehat{\mathbf{f}}_k(X_k) = \mathbf{\Psi}_k(\mathbf{\Psi}_k^{\mathrm{T}} \mathbf{\Psi}_k)^{-1} \mathbf{\Psi}_k^{\mathrm{T}} \mathbf{Y}.$$

We can treat

$$\widehat{\omega}_k^{\mathrm{mNIS}} = \|\widehat{\mathbf{f}}_k(X_k)\|_{\mathrm{F}}^2 \tag{2.5}$$

as the marginal utility of the multi-response nonparametric independence screening (mNIS) for X_k. Equivalently, we can rank the predictors in descending order of the RSS of the marginal nonparametric regressions (2.4). NIS can fail if the underlying additive regression model assumption fails. It is expected that mNIS inherits this drawback of NIS.

2.3. Robust rank correlation screening

Li et al. (2012) proposed using Kendall's τ correlation coefficient as a ranking statistic. Their method is named RRCS. The marginal utility they propose is equal to a quarter of Kendall's τ correlation coefficient; that is,

$$\frac{1}{4}\widehat{\tau}(X_k, Y) = \frac{1}{n(n-1)} \sum_{i \neq l}^{n} I(X_{ik} < X_{lk}) I(Y_i < Y_l) - \frac{1}{4}.$$

Similarly to (2.3), we try to extend this to the multiple-response case by simply summing up the squares of the Kendall's τ correlations between the predictor and each response,

$$\widehat{\omega}_k^{\mathrm{mRRCS}} = \|\widehat{\boldsymbol{\tau}}_k\|^2, \tag{2.6}$$

where $\widehat{\boldsymbol{\tau}}_k = \left(\widehat{\tau}(X_k, Y_1), \ldots, \widehat{\tau}(X_k, Y_q)\right)^{\mathrm{T}}$. We refer to (2.6) as the multi-response robust rank correlation screening (mRRCS) statistic. The population version of Kendall's τ correlation is zero if two random variables are independent; as a result, $\omega_k^{\mathrm{mRRCS}}$ is zero if X_k is independent of the multivariate responses.

2.4. Distance correlation screening

The distance correlation (DC) (Székely, Rizzo and Bakirov (2007)) measures the dependence between two random vectors. Unlike the Pearson correlation and Kendall's τ correlation, the DC is equal to zero if and only if the two random vectors are independent. This unique property motivated Li, Zhong and Zhu (2012) to consider distance correlation screening (DCS), which has become one of the most popular model-free variable screening methods. DC and DCS can be applied naturally to multi-response data. For the sake of completeness, we briefly review DCS here. The DC can be computed using the distance covariance. For a given sample $\{\mathbf{U}_i, \mathbf{V}_i\}_{i=1}^n$ from two random vectors $\mathbf{U}, \mathbf{V}$, the squared distance covariance can be estimated as

$$\widehat{\mathrm{dcov}}^2(\mathbf{U}, \mathbf{V}) = \widehat{S}_1(\mathbf{U}, \mathbf{V}) + \widehat{S}_2(\mathbf{U}, \mathbf{V}) - 2\widehat{S}_3(\mathbf{U}, \mathbf{V}),$$

where

$$\begin{aligned}
\widehat{S}_1(\mathbf{U}, \mathbf{V}) &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\mathbf{U}_i - \mathbf{U}_j\| \, \|\mathbf{V}_i - \mathbf{V}_j\|, \\
\widehat{S}_2(\mathbf{U}, \mathbf{V}) &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\mathbf{U}_i - \mathbf{U}_j\| \, \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\mathbf{V}_i - \mathbf{V}_j\|, \\
\widehat{S}_3(\mathbf{U}, \mathbf{V}) &= \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n \|\mathbf{U}_i - \mathbf{U}_l\| \, \|\mathbf{V}_j - \mathbf{V}_l\|.
\end{aligned}$$

Therefore, the DCS can be implemented by computing

$$\widehat{\omega}_k^{\mathrm{DCS}} = \widehat{\mathrm{dcor}}(X_k, \mathbf{Y}) = \frac{\widehat{\mathrm{dcov}}(X_k, \mathbf{Y})}{\sqrt{\widehat{\mathrm{dcov}}(X_k, X_k)\widehat{\mathrm{dcov}}(\mathbf{Y}, \mathbf{Y})}}$$

for each predictor X_k.

Numerical studies (Li, Zhong and Zhu (2012)) have shown DCS exhibits good performances for very complex models. In general, DCS outperforms SIS, unless the true model is a linear regression model. Li, Zhong and Zhu (2012) proved that the sure screening property of DCS holds if the responses and predictors are sub-Gaussian. On the other hand, DCS performs poorly in the presence of heavy-tailed data (Mai and Zou (2015)). Hence, sub-Gaussian tail assumptions seem to be necessary and sufficient for the sure screening property of DCS. Another important drawback of DCS is that it is not invariant against monotonic transformation, unlike, for example, rank correlation screening (Li et al. (2012)) and the fused Kolmogorov filter (Mai and Zou (2015)).

3. A New Approach: Rank Canonical Correlation Screening

In this section, we first review a useful tool in multivariate analysis called the canonical correlation analysis (CCA), and then introduce a novel way to combine a rank correlation and a canonical correlation.

3.1. Canonical correlation

A canonical correlation analysis is a way of inferring information from cross-covariance matrices by finding two projections for two random vectors, such that the projected random vectors have maximum correlation with each other. For the kth predictor X_k, the canonical correlation between X_k and $(Y_1, \ldots, Y_q)$ is defined as

$$\rho_k^c = \max_{\mathbf{b}} \frac{\boldsymbol{\Sigma}^*_{X_k \mathbf{Y}} \mathbf{b}}{\sqrt{\Sigma^*_{X_k}} \sqrt{\mathbf{b}^{\mathrm{T}} \boldsymbol{\Sigma}^*_{\mathbf{Y}} \mathbf{b}}},$$

where $\Sigma^*_{X_k}, \boldsymbol{\Sigma}^*_{X_k \mathbf{Y}}$, and $\boldsymbol{\Sigma}^*_{\mathbf{Y}}$ are submatrices of $\boldsymbol{\Sigma}^* = \begin{pmatrix} \Sigma^*_{X_k} & \boldsymbol{\Sigma}^*_{X_k \mathbf{Y}} \\ \boldsymbol{\Sigma}^*_{\mathbf{Y} X_k} & \boldsymbol{\Sigma}^*_{\mathbf{Y}} \end{pmatrix}$, which is the covariance matrix of $\big(X_k, (Y_1, \ldots, Y_q)\big)$. By the Cauchy–Schwartz inequality, it can be shown that

$$\rho_k^c = \frac{\left(\boldsymbol{\Sigma}^*_{X_k \mathbf{Y}} {\boldsymbol{\Sigma}^*_{\mathbf{Y}}}^{-1} \boldsymbol{\Sigma}^*_{\mathbf{Y} X_k}\right)^{1/2}}{\sqrt{\Sigma^*_{X_k}}}. \tag{3.1}$$

The canonical correlation can also be related to Pearson correlations. Note that the square of (3.1) is equivalent to

$$(\rho_k^c)^2 = \boldsymbol{\rho}_k^{\mathrm{T}} (\boldsymbol{\Sigma}_{\mathbf{Y}})^{-1} \boldsymbol{\rho}_k, \tag{3.2}$$

recalling that $\boldsymbol{\rho}_k = (\rho_{k1}, \ldots, \rho_{kq})^{\mathrm{T}}$ are the Pearson correlations between X_k and Y_j, and $\boldsymbol{\Sigma}_{\mathbf{Y}} = (\rho_{jl})_{q \times q}$ is the correlation matrix of $(Y_1, \ldots, Y_q)$. We can use

$$\widehat{\omega}_k^{\mathrm{mCC}} = (\widehat{\rho}_k^c)^2 = \widehat{\boldsymbol{\rho}}_k^{\mathrm{T}} (\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}})^{-1} \widehat{\boldsymbol{\rho}}_k \tag{3.3}$$

as a variable screening statistic, which we refer to as the multi-response canonical correlation screening (mCC) statistic.

To compare mCC and mSIS, we see that ω_k^{mSIS} simply sums up ρ_{kj}^2, whereas (3.2) is a weighted summation of ρ_{kj}^2 and $\rho_{kj}\rho_{kl}$, $(j \neq l)$, recruiting the information on the cross-correlation among Y_j. In other words, mCC is able to use the joint information of the multiple responses and the covariate.

The mCC statistic is still a linear correlation measure. We would like to consider a generalization that can capture the nonlinear correlation between the response vector and the predictor. We introduce such a method in the next subsection.

3.2. Rank canonical correlation screening

Inspired by the robust advantage of rank correlation, we consider a better version of mCC that integrates rank correlation with canonical correlation. Two commonly used rank correlations, Kendall's τ correlation and Spearman's rank correlation, are employed.

We first replace the Pearson correlations between X_k and Y_j in (3.3) with the corresponding Spearman rank correlations,

$$(\widehat{r}_k^c)^2 \triangleq \widehat{\mathbf{r}}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}^{-1} \widehat{\mathbf{r}}_k, \tag{3.4}$$

where $\widehat{\mathbf{r}}_k = (\widehat{r}_s(X_k, Y_1), \ldots, \widehat{r}_s(X_k, Y_q))^{\mathrm{T}}$ are the Spearman correlation coefficients between X_k and Y_j, and $\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})} = (\widehat{r}_s(Y_j, Y_l))_{q\times q}$ is a matrix of Spearman correlations between all pairs of Y_j and Y_l. The Spearman rank correlation (Spearman (1904); Durbin and Stuart (1951)) that measures an ordinal association is analogous to the Pearson correlation between the rank values of two variables. Thus, $\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}$ is exactly the sample correlation matrix of $\big(R(Y_1), \ldots, R(Y_q)\big)$, where $R(\cdot)$ stands for the rank of a random variable among n observations.

Next, we adopt Kendall's τ correlation in (3.4), and define

$$(\widehat{\tau}_k^c)^2 \triangleq \widehat{\boldsymbol{\tau}}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\widetilde{R}(\mathbf{Y})}^{-1} \widehat{\boldsymbol{\tau}}_k, \tag{3.5}$$

recalling that $\widehat{\boldsymbol{\tau}}_k = (\widehat{\tau}(X_k, Y_1), \ldots, \widehat{\tau}(X_k, Y_q))^{\mathrm{T}}$ are the Kendall τ correlations between X_k and Y_j, and $\widehat{\boldsymbol{\Sigma}}_{\widetilde{R}(\mathbf{Y})} = (\widehat{\tau}(Y_j, Y_l))_{q\times q}$ is a matrix of Kendall τ correlations between all pairs of Y_j and Y_l.

Therefore, we propose a screening approach based on (3.4) or (3.5) as the multi-response rank canonical correlation screening (mRCC) statistic, and denote the ranking measures as $\widehat{\omega}_k^{\mathrm{mRCC1}}$ or $\widehat{\omega}_k^{\mathrm{mRCC2}}$, respectively. With a prespecified threshold t_n or t'_n, we select the set

$$\widehat{\mathcal{A}}_{t_n} = \{1 \le k \le p : \widehat{\omega}_k^{\mathrm{mRCC1}} \ge t_n\} \;\text{ or }\; \widehat{\mathcal{A}'}_{t'_n} = \{1 \le k \le p : \widehat{\omega}_k^{\mathrm{mRCC2}} \ge t'_n\}$$

as the important variables, respectively. In Section 4, we establish the sure screening properties of mRCC1 and mRCC2. In practice, we can also pick the top

d_n variables with the top d_n mRCC1 or mRCC2 values, where $d_n = c[n/\log n]$, for $c = 1$ or 2.

Remark 1. The proposed mRCC not only inherits the multivariate merits of the canonical correlation, taking advantage of the joint information of the multiple responses and the covariate, but also preserves nice properties of the rank correlation that can handle heavy-tailed predictors and responses, as well as invariance against monotonic transformations of them. Because of these nice properties and its excellent numerical performance in Section 5, mRCC is the main method we advocate using in practice.

4. The Sure Screening Property

We establish the sure screening property of mRCC1 and mRCC2 in this section. Following (Li, Zhong and Zhu (2012)), we define the true predictor subset as

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

with size $s = |\mathcal{A}|$. For variable X_k, the population versions of mRCC1 and mRCC2 are

$$\omega_k^{\mathrm{mRCC1}} = (r_k^c)^2 = \mathbf{r}_k^{\mathrm{T}} \boldsymbol{\Sigma}_{R(\mathbf{Y})}^{-1} \mathbf{r}_k \tag{4.1}$$

and

$$\omega_k^{\mathrm{mRCC2}} = (\tau_k^c)^2 = \boldsymbol{\tau}_k^{\mathrm{T}} \boldsymbol{\Sigma}_{\widetilde{R}(\mathbf{Y})}^{-1} \boldsymbol{\tau}_k, \tag{4.2}$$

respectively, where $\mathbf{r}_k = (r_s(X_k, Y_1), \ldots, r_s(X_k, Y_q))^{\mathrm{T}}$, $\boldsymbol{\Sigma}_{R(\mathbf{Y})} = (r_s(Y_j, Y_l))_{q\times q}$, and $\boldsymbol{\tau}_k = (\tau(X_k, Y_1), \ldots, \tau(X_k, Y_q))^{\mathrm{T}}$, $\boldsymbol{\Sigma}_{\widetilde{R}(\mathbf{Y})} = (\tau(Y_j, Y_l))_{q\times q}$. For two random variables U and V from a joint distribution, let $(U_1, V_1), (U_2, V_2)$, and (U_3, V_3) be three independent realizations. Then, $r_s(U, V) = \mathrm{cov}\big(\mathrm{sgn}(U_1 - U_2), \mathrm{sgn}(V_1 - V_3)\big)$ and $\tau(U, V) = \mathrm{cov}\big(\mathrm{sgn}(U_1 - U_2), \mathrm{sgn}(V_1 - V_2)\big)$.

We consider the following conditions:

(C1) There exists a positive c_0 such that $\lambda_{\min}\big(\boldsymbol{\Sigma}_{R(\mathbf{Y})}\big) \geq c_0 q^{-1}$;

(C2) There exists $\widetilde{A}$, a subset of $\{1, \ldots, p\}$, and a constant $0 < \kappa < 1/2$, such that $|\widetilde{A}| \leq |\widehat{\mathcal{A}}_{t_n}|$, $\mathcal{A} \subset \widetilde{A}$, and $\delta_{\widetilde{\mathcal{A}}} = q^{-4} n^{\kappa} \{\min_{k\in\widetilde{\mathcal{A}}} \omega_k^{\mathrm{mRCC1}} - \max_{k\in\widetilde{\mathcal{A}}^c} \omega_k^{\mathrm{mRCC1}}\} > 0$;

(C1′) There exists a positive c_0' such that $\lambda_{\min}\big(\boldsymbol{\Sigma}_{\widetilde{R}(\mathbf{Y})}\big) \geq c_0' q^{-1}$;

(C2′) There exists $\widetilde{A'}$, a subset of $\{1, \ldots, p\}$, and a constant $0 < \kappa < 1/2$, such that $|\widetilde{A'}| \le |\widehat{\mathcal{A}'}_{t'_n}|$, $\mathcal{A} \subset \widetilde{A'}$ and $\delta_{\widetilde{\mathcal{A}'}} = q^{-4}n^{\kappa}\{\min_{k\in\widetilde{\mathcal{A}'}}\ \omega_k^{\mathrm{mRCC2}} - \max_{k\in\widetilde{\mathcal{A}'}^c}\ \omega_k^{\mathrm{mRCC2}}\} > 0$.

Conditions (C1) and (C1′) rule out the cases in which one component of the multivariate responses is a perfect monotonic function of another with a perfect Spearman correlation of $+1$ or -1, and that, the agreement or the disagreement between the rankings of two components of the response is perfect with a perfect Kendall's τ correlation of $+1$ or -1, respectively. Conditions (C2) and (C2′) are very common in the screening literature (Mai and Zou (2013, 2015)), and are the theoretical basis of the sure screening property. They assume there is a gap between the marginal signals inside and outside a subset containing the true predictor subset.

The following theorem gives the sure screening property of mRCC.

Theorem 1.

1. *Under Condition* (C1)*, for any* $c_2 > 0$ *and* $0 < \kappa < 1/2$*, there exist some positive constants* c_3, c_4, c_5, c_6*, and* C*, such that when* $n > \max\{Cq^2, 6^{1/(1-\kappa)}\}$,

$$\begin{aligned}&\Pr\Big(\max_{1\le k\le p}|\widehat{\omega}_k^{\mathrm{mRCC1}} - \omega_k^{\mathrm{mRCC1}}| \ge c_2 q^4 n^{-\kappa}\Big)\\&\le p\cdot\Big\{6q^2\Big(\exp(-c_3 n q^{-4}) + \exp(-c_4 n^3 q^{-4})\Big)\\&\quad + (2q^2+4q)\Big(\exp(-c_5 n^{1-2\kappa}) + \exp(-c_6 n^{3-2\kappa})\Big)\Big\}.\end{aligned}$$

Under Condition (C1′)*, for any* $c'_2 > 0$ *and* $0 < \kappa < 1/2$*, there exist some positive constants* c'_3, c'_4*, such that*

$$\begin{aligned}&\Pr\Big(\max_{1\le k\le p}|\widehat{\omega}_k^{\mathrm{mRCC2}} - \omega_k^{\mathrm{mRCC2}}| \ge c'_2 q^4 n^{-\kappa}\Big)\\&\le p\cdot\Big\{6q^2\exp(-c'_3 n q^{-4}) + (2q^2+4q)\exp(-c'_4 n^{1-2\kappa})\Big\}.\end{aligned}$$

2. *If Conditions* (C1) *and* (C2) *hold and we set* $t_n = c_1 q^4 n^{-\kappa}$*, with* $c_1 \le \delta_{\widetilde{\mathcal{A}}}/2$*, we have*

$$\begin{aligned}\Pr\big(\mathcal{A}\subset\widehat{\mathcal{A}}_{t_n}\big) \ge 1 - p\cdot\Big\{&6q^2\Big(\exp(-c_3 n q^{-4}) + \exp(-c_4 n^3 q^{-4})\Big)\\&+(2q^2+4q)\Big(\exp(-c_5 n^{1-2\kappa}) + \exp(-c_6 n^{3-2\kappa})\Big)\Big\}.\end{aligned}$$

If Conditions (C1′) *and* (C2′) *hold and we set* $t'_n = c'_1 q^4 n^{-\kappa}$ *with* $c'_1 \le \delta_{\widetilde{\mathcal{A}'}}/2$, *we have*

$$\Pr\left(\mathcal{A} \subset \widehat{\mathcal{A}'}_{t_n'}\right) \geq 1 - p \cdot \Big\{6q^2 \exp(-c_3' n q^{-4}) + (2q^2+4q)\exp(-c_4' n^{1-2\kappa})\Big\}.$$

Theorem 1 gives an upper bound on the dimension of the response, $q = o(n^{1/4})$, to have the sure screening property. It also follows from Theorem 1 that the limit of the data dimensionality we can handle should satisfy $\log(pq^2) = o(nq^{-4} + n^{1-2\kappa})$ using both methods, with $0 < \kappa < 1/2$. Under these settings we have the sure screening properties $\Pr\left(\mathcal{A} \subset \widehat{\mathcal{A}}_{t_n}\right) \to 1$ and $\Pr\left(\mathcal{A} \subset \widehat{\mathcal{A}'}_{t_n'}\right) \to 1$, respectively.

In contrast to the sub-Gaussian distribution assumption required for the sure screening property of DCS (Li, Zhong and Zhu (2012)), we do not require any assumption on the moments of the predictors or the responses for the sure screening property of mRCC.

Theorem 2. *Under Condition* (C1)*, for any* $t_n = c_1 q^4 n^{-\kappa}$*, there exist some positive constants* c_3*,* c_4*,* c_5*,* c_6*, and* C*, such that when* $n > \max\{Cq^2, 6^{1/(1-\kappa)}\}$*,*

$$\Pr\left(|\widehat{\mathcal{A}}_{t_n}| \leq O(sq^{-2}n^{\kappa})\right) \geq 1 - p \cdot \Big\{6q^2\Big(\exp(-c_3 n q^{-4}) + \exp(-c_4 n^3 q^{-4})\Big) + (2q^2+4q)\Big(\exp(-c_5 n^{1-2\kappa}) + \exp(-c_6 n^{3-2\kappa})\Big)\Big\}.$$

Under Condition (C1′)*, for any* $t_n' = c_1' q^4 n^{-\kappa}$*, there exist some positive constants* c_3'*,* c_4'*, such that*

$$\Pr\left(|\widehat{\mathcal{A}'}_{t_n'}| \leq O(sq^{-2}n^{\kappa})\right) \geq 1 - p \cdot \Big\{6q^2 \exp(-c_3' n q^{-4}) + (2q^2+4q)\exp(-c_4' n^{1-2\kappa})\Big\}.$$

This result controls the model size of the selected model, which is of order $O(sq^{-2}n^{\kappa})$. The false selection rate converges to zero exponentially fast.

5. Numerical Studies

In this section, we evaluate the performance of all screening procedures discussed in this paper using simulation experiments and a real-data analysis. As suggested by a referee, we include a newly published variable screening method called composite coefficient of determination (CCD), proposed by Kong, Xia and Zhong (2019), and extend it to the multiple responses case in a similar way to mSIS by aggregating the ranking statistics of the predictor with each response. We denote this method as mCCD. For the derivation and explanations of CCD, refer to Kong, Xia and Zhong (2019).

5.1. Simulations

We repeat each simulation 200 times, and use the following three criteria, adopted by Li, Zhong and Zhu (2012):

1. $\mathcal{S}$: the minimum model size that includes all true predictors. We report the 5%, 25%, 50%, 75%, and 95% quantiles of $\mathcal{S}$ out of 200 replications.

2. $\mathcal{P}_s$: the proportion that an individual true predictor is selected for a given model size d in the 200 replications.

3. $\mathcal{P}_a$: the proportion that all true predictors are selected for a given model size d in the 200 replications.

Here, $\mathcal{S}$ measures the accuracy of a screening procedure. The smaller $\mathcal{S}$ is, the less complex the resulting model is, and the better the screening procedure is. Then, $\mathcal{P}_s$ and $\mathcal{P}_a$ allow us to examine the chance that a screening procedure misses an individual predictor and all true predictors, respectively, for a given model size d. We present the simulation results of $\mathcal{P}_s, \mathcal{P}_a$ with $d = 2[n/\log n]$ for all the examples and the real data. We also tried $d = [n/\log n]$, with similar outcomes; hence, we omit such results here for brevity.

Example 1. We adopt the simple linear model from Fan and Lv (2008):

$$Y_j = 5X_1 + 5X_2 + 5X_3 + \epsilon_j, \quad j = 1, 2, \ldots, 10. \tag{5.1}$$

The predictor vector $(X_1, \ldots, X_p)$ is drawn from a multivariate normal distribution $N(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = CS(\rho)$ is a compound symmetric matrix, with all entries being ρ, except for the diagonal elements being one, and the noise ϵ_j follows a standard normal distribution. The sample size is $n = 50$, the numbers of the predictors and responses are $p = 1{,}000$, and $q = 10$, respectively, and we consider $\rho = 0, 0.1, 0.5, 0.9$.

Table 1 summarizes the simulation results for $\mathcal{S}$, $\mathcal{P}_s$, and $\mathcal{P}_a$. We can see that the mSIS works best in this example because this model is actually a univariate response data linear model with a strong signal-to-noise ratio in every component.For the mCC and mRCC1, the performance is acceptable, albeit a little worse than that of the other methods. One reason is that each response has the same strong signal, which dominates the error term. Hence, the Pearson correlations between the pairs of responses are almost one(about 0.98 and 0.99). In such a case, the Condition (C1) for mRCC1 may be violated. This may also be true for the mCC because there is an inverse correlation matrix of the responses

Table 1. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ and the proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ out of 200 replications in Example 1.

		$\mathcal{S}$					$\mathcal{P}_s$			$\mathcal{P}_a$
ρ	Method	5%	25%	50%	75%	95%	X1	X2	X3	All
0	mSIS	3.0	3.0	3.0	5.0	16.1	0.98	0.98	0.94	0.90
	mNIS	3.0	4.0	8.0	30.3	144.8	0.92	0.98	0.93	0.89
	mRRCS	3.0	3.0	4.0	7.0	33.2	0.97	0.96	0.90	0.74
	DCS	3.0	3.0	4.0	7.0	27.1	0.98	0.97	0.90	0.85
	mCCD	3.0	3.0	4.0	7.0	24.2	0.98	0.97	0.87	0.57
	mCC	4.0	11.0	25.0	100.8	283.2	0.83	0.85	0.88	0.84
	mRCC1	4.0	15.8	39.5	130.5	427.0	0.75	0.81	0.77	0.60
	mRCC2	3.0	3.0	4.0	9.0	40.1	0.97	0.96	0.91	0.72
0.1	mSIS	3.0	3.0	3.0	5.0	23.1	1.00	1.00	0.92	0.90
	mNIS	3.0	3.0	6.0	12.0	41.4	0.86	0.96	0.87	0.87
	mRRCS	3.0	3.0	4.0	8.0	41.2	0.97	0.98	0.84	0.71
	DCS	3.0	3.0	4.0	7.0	35.1	0.98	0.99	0.89	0.81
	mCCD	3.0	3.0	4.0	7.0	34.0	0.99	0.99	0.85	0.57
	mCC	3.0	6.0	16.0	51.3	144.6	0.76	0.83	0.80	0.82
	mRCC1	4.0	9.0	24.5	67.0	226.9	0.74	0.80	0.68	0.60
	mRCC2	3.0	3.0	4.0	9.0	44.1	0.96	0.99	0.83	0.71
0.5	mSIS	3.0	4.0	6.5	18.3	100.4	0.99	0.98	0.93	0.89
	mNIS	3.0	5.0	9.0	27.0	124.2	0.92	0.97	0.91	0.89
	mRRCS	3.0	6.0	13.0	35.3	175.4	0.98	0.97	0.88	0.77
	DCS	3.0	5.0	11.0	29.0	141.2	0.98	0.97	0.90	0.88
	mCCD	3.0	7.0	14.5	40.5	192.7	0.99	0.97	0.87	0.60
	mCC	3.0	6.0	14.0	41.0	208.5	0.78	0.86	0.88	0.88
	mRCC1	4.0	11.8	38.0	83.0	371.1	0.68	0.83	0.74	0.65
	mRCC2	3.0	6.0	14.0	35.3	186.6	0.98	0.97	0.87	0.76
0.9	mSIS	3.0	4.0	8.0	19.8	91.1	0.96	0.95	0.80	0.76
	mNIS	3.0	5.0	11.0	28.0	105.1	0.71	0.91	0.74	0.72
	mRRCS	4.0	13.0	36.0	77.5	231.0	0.91	0.91	0.67	0.40
	DCS	3.0	6.0	16.0	39.3	142.5	0.94	0.93	0.71	0.62
	mCCD	5.0	34.8	75.0	173.5	424.4	0.95	0.93	0.65	0.20
	mCC	3.0	7.0	14.5	38.3	166.3	0.50	0.61	0.61	0.64
	mRCC1	8.0	25.0	65.0	144.5	437.5	0.36	0.50	0.39	0.25
	mRCC2	4.0	13.0	36.0	77.0	258.7	0.91	0.92	0.67	0.40

in (3.2). Another reason is that the sample size is not big enough; therefore, the rank-based mRCC1 may lose some efficiency.

Example 2. Consider the following generalized Box–Cox transformation model adapted from Li et al. (2012):

$$H(Y_j) = X_{10j-9} + X_{10j-8} + \epsilon_j, \quad j = 1, 2, \ldots, 10, \tag{5.2}$$

where the transformation functions are unknown. In the simulation, we consider the Box–Cox transformation:

$$H(Y) = \frac{|Y|^{\lambda}\mathrm{sgn}(Y) - 1}{\lambda}, \text{when } \lambda = 0.25, 0.5, 0.75, 1; \quad H(Y) = \log Y, \text{when } \lambda = 0.$$

The variables $(X_1, \ldots, X_p)$ and the noise ϵ_j are generated in the same way as in Example 1. The number of true variables is 20, and $(n, p, q) = (200, 2000, 10)$ and $\rho = 0.1, 0.5$.

The simulation results for $\mathcal{S}$ and $\mathcal{P}_a$ are reported in Tables 2 and 3, respectively. We can see clearly that mRCC1 significantly outperforms the other methods, especially when $\rho = 0.5$, and the results are almost invariant under transformations (there are small differences owing to the different random errors generated for models with different λ). Although mRRCS is also rank-based and invariant under transformation, it performs poorly when $\rho = 0.5$. The reason may be that it ignores the dependence structure of the multivariate responses. When the model deviates from a linear model (λ decrease from one), the performance of mSIS, mNIS, DCS, and mCC quickly deteriorates, owing to the existence of the nonlinearity and the heavy-tailed responses.

Example 3. In this example, we consider the following model:

$$Y_j = 2\sin\left(\alpha_{j1}X_1 + \alpha_{j2}X_2 + \alpha_{j3}X_3 + \alpha_{j4}X_4 + \alpha_{j5}X_5\right) + \epsilon_j, \ j = 1, 2, \ldots, 20, \tag{5.3}$$

where $\alpha_{j1}, \ldots, \alpha_{j5} \sim$ Unif(0,1) independently, for $j = 1, \ldots, 20$. Once the parameter is drawn, the model is fixed. We generate $(X_1, \ldots, X_p)$ and the noise ϵ_j as in Example 1, and $(n, p, q) = (200, 2000, 20)$ and $\rho = 0.5, 0.8$.

Table 4 gives the results for $\mathcal{S}$. Table 5 shows the results for $\mathcal{P}_s$ and $\mathcal{P}_a$. For this nonlinear model, mRCC1 and mRCC2 are still robust and encouraging, and perform best.

Example 4. We adopt the additive model from Mai and Zou (2015):

$$Y_j = 4X_{jk_1} + 2\tan\left(\frac{\pi X_{jk_2}}{2}\right) + 5X_{jk_3}^2 + \epsilon_j, \quad j = 1, 2, \ldots, 20. \tag{5.4}$$

The predictors follow Unif(0,1) independently, and ϵ_j follow $N(0, 1)$, and are independent of the predictors. For each j, the indices $\{k_1, k_2, k_3\}$ are drawn randomly from $\{1, 2, \ldots, 10\}$. Once the indices are drawn, the model is fixed. In our simulation, we check that $X_1, X_2, \ldots, X_{10}$ are all included in the model;

Table 2. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ out of 200 replications in Example 2.

		CS(0.1)					CS(0.5)				
λ	Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0	mSIS	202.7	421.0	653.0	940.3	1,413.7	1,237.9	1,553.8	1,723.5	1,829.0	1,945.3
	mNIS	600.6	880.5	1,118.5	1,310.5	1,655.4	1,296.9	1,563.3	1,729.0	1,876.3	1,955.2
	mRRCS	20.0	23.0	28.5	43.0	76.3	745.4	1,054.5	1,336.5	1,547.5	1,816.5
	DCS	423.4	712.8	924.5	1,194.0	1,665.2	1,143.0	1,485.3	1,685.0	1,825.0	1,957.3
	mCCD	25.0	39.0	63.0	102.0	263.9	905.0	1,251.5	1,440.5	1,653.5	1,861.4
	mCC	137.0	294.3	549.0	869.3	1,440.7	1,057.0	1,346.5	1,563.0	1,748.8	1,924.2
	mRCC1	20.0	20.0	20.0	20.0	21.1	40.0	96.0	155.5	313.0	731.2
	mRCC2	20.0	20.0	20.0	20.0	24.0	95.9	179.0	288.5	513.8	1,012.1
0.25	mSIS	30.0	46.0	71.5	138.3	397.1	1,128.0	1,363.0	1,598.0	1,767.3	1,923.7
	mNIS	141.0	254.0	346.0	485.0	827.5	1,078.9	1,432.3	1,606.0	1,774.3	1,919.3
	mRRCS	20.0	23.0	28.0	43.3	101.2	733.9	1,027.0	1,277.0	1,542.8	1,786.0
	DCS	56.0	103.8	168.0	268.8	601.3	981.8	1,305.5	1,490.5	1,696.3	1,925.5
	mCCD	21.0	27.0	36.0	59.3	158.0	966.5	1,183.5	1,400.0	1,631.8	1,890.0
	mCC	20.0	21.0	24.0	34.3	84.1	360.9	591.8	827.0	1,108.3	1,621.0
	mRCC1	20.0	20.0	20.0	20.0	22.0	45.8	90.5	151.0	292.0	693.3
	mRCC2	20.0	20.0	20.0	20.0	24.0	92.0	192.5	304.5	485.0	1,034.1
0.5	mSIS	22.0	27.8	41.0	65.8	173.0	942.0	1,214.5	1,453.5	1,654.0	1,888.2
	mNIS	59.0	96.5	151.0	237.0	503.3	1,056.0	1,310.0	1,498.0	1,686.8	1,885.1
	mRRCS	20.0	22.0	29.0	43.0	105.1	720.6	1,044.8	1,332.5	1,509.5	1,826.1
	DCS	25.0	36.8	54.0	99.0	235.2	884.2	1,191.3	1,409.0	1,606.3	1,886.6
	mCCD	22.0	25.0	36.0	58.0	129.4	901.6	1,172.3	1,415.0	1,622.5	1,889.5
	mCC	20.0	20.0	20.0	21.0	28.0	90.6	170.5	317.0	504.0	995.3
	mRCC1	20.0	20.0	20.0	20.0	22.0	45.0	87.8	175.0	310.5	832.1
	mRCC2	20.0	20.0	20.0	20.0	23.1	84.8	197.3	330.0	502.5	1,097.3
0.75	mSIS	21.0	25.0	33.0	51.0	106.6	795.0	1,122.0	1,390.0	1,625.5	1,879.0
	mNIS	27.0	43.0	66.5	118.3	301.5	845.6	1,132.0	1,420.0	1,613.5	1,859.7
	mRRCS	21.0	23.8	29.0	43.0	100.5	740.8	996.5	1,273.0	1,547.3	1,822.3
	DCS	21.0	25.0	32.5	50.3	117.1	766.7	1,079.5	1,325.0	1,593.8	1,821.3
	mCCD	21.0	27.0	36.0	57.0	130.1	900.2	1,162.0	1,380.5	1,649.3	1,872.4
	mCC	20.0	20.0	20.0	20.0	21.0	32.0	55.8	109.5	216.0	495.2
	mRCC1	20.0	20.0	20.0	20.0	23.0	41.0	98.5	163.0	303.5	613.6
	mRCC2	20.0	20.0	20.0	20.3	27.0	87.0	200.8	317.0	529.3	1,008.1
1	mSIS	21.0	24.0	29.5	43.0	89.7	661.9	1,126.0	1,358.0	1,616.8	1,841.0
	mNIS	24.0	38.0	59.0	109.3	228.2	686.7	1,170.0	1,391.0	1,596.5	1,862.1
	mRRCS	20.0	22.8	28.0	42.0	99.1	649.1	1,064.5	1,322.0	1,542.8	1,812.1
	DCS	21.0	24.0	30.5	44.0	95.1	702.6	1,099.5	1,351.5	1,576.0	1,809.6
	mCCD	22.0	26.0	34.0	56.5	120.4	776.7	1,174.0	1,427.5	1,651.8	1,845.1
	mCC	20.0	20.0	20.0	20.0	21.0	27.0	50.5	84.5	152.8	473.8
	mRCC1	20.0	20.0	20.0	20.0	22.1	47.9	86.5	157.5	308.5	754.3
	mRCC2	20.0	20.0	20.0	20.0	25.0	92.4	189.0	319.0	533.3	1,144.2

Table 3. The proportions of $\mathcal{P}_a$ in Example 2.

	CS(0.1)					CS(0.5)				
Method	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$
mSIS	0.00	0.52	0.80	0.90	0.92	0.00	0.00	0.00	0.00	0.00
mNIS	0.00	0.00	0.14	0.57	0.59	0.00	0.00	0.00	0.00	0.00
mRRCS	0.95	0.94	0.90	0.91	0.89	0.00	0.00	0.00	0.00	0.00
DCS	0.00	0.14	0.63	0.87	0.91	0.00	0.00	0.00	0.00	0.00
mCCD	0.58	0.83	0.84	0.84	0.85	0.00	0.00	0.00	0.00	0.00
mCC	0.02	0.94	1.00	1.00	1.00	0.00	0.00	0.03	0.36	0.41
mRCC1	1.00	1.00	1.00	1.00	1.00	0.18	0.17	0.18	0.15	0.18
mRCC2	1.00	1.00	1.00	1.00	1.00	0.03	0.03	0.03	0.02	0.04

Table 4. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ out of 200 replications in Example 3.

	CS(0.5)					CS(0.8)				
Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
mSIS	5.0	5.0	9.5	33.2	211.0	22.9	104.8	303.5	573.0	1,392.2
mNIS	5.0	5.0	6.0	10.0	83.4	5.0	11.0	32.0	123.2	497.7
mRRCS	5.0	5.0	6.0	10.2	71.1	7.0	18.0	59.5	179.5	592.2
DCS	5.0	5.0	6.0	9.0	71.0	6.0	15.8	62.5	172.2	677.9
mCCD	5.0	5.0	7.0	17.0	96.7	11.0	41.0	125.5	301.2	928.8
mCC	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	16.1
mRCC1	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
mRCC2	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.0

Table 5. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 3.

	CS(0.5)						CS(0.8)					
	$\mathcal{P}_s$					$\mathcal{P}_a$	$\mathcal{P}_s$					$\mathcal{P}_a$
Method	X1	X2	X3	X4	X5	All	X1	X2	X3	X4	X5	All
mSIS	0.86	1.00	1.00	1.00	1.00	0.86	0.27	0.95	0.83	0.94	0.82	0.17
mNIS	0.93	1.00	1.00	1.00	1.00	0.93	0.69	0.99	0.99	1.00	1.00	0.66
mRRCS	0.96	1.00	1.00	1.00	1.00	0.96	0.57	0.99	0.98	0.99	1.00	0.54
DCS	0.96	1.00	1.00	1.00	1.00	0.96	0.59	0.98	0.97	1.00	1.00	0.54
mCCD	0.92	1.00	1.00	1.00	1.00	0.92	0.46	0.97	0.96	1.00	0.99	0.42
mCC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
mRCC1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
mRCC2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

hence, the number of true predictors is 10. The sample size is $n = 200$, and the numbers of predictors and responses are $p = 2{,}000$ and $q = 20$, respectively.

From Table 6 and Table 7, we see that the mRCC1 and mRCC2 achieve

Table 6. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ out of 200 replications in Example 4.

Method	5%	25%	50%	75%	95%
mSIS	908.2	1,335.0	1,573.0	1,794.3	1,960.1
mNIS	1,085.4	1,479.8	1,665.5	1,835.5	1,965.2
mRRCS	10.0	11.0	13.0	25.0	76.0
DCS	275.6	600.0	942.0	1,375.5	1,806.4
mCCD	15.0	56.8	179.5	368.5	807.4
mCC	10.0	10.0	17.0	141.3	912.1
mRCC1	10.0	10.0	10.0	10.0	10.0
mRCC2	10.0	10.0	10.0	10.0	10.0

Table 7. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 4.

	$\mathcal{P}_s$										$\mathcal{P}_a$
Method	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	All
mSIS	0.10	0.03	0.90	0.09	0.07	0.92	0.07	0.78	0.97	0.88	0.00
mNIS	0.04	0.02	0.99	0.05	0.04	1.00	0.02	0.92	1.00	1.00	0.00
mRRCS	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95
DCS	0.49	0.16	0.91	0.42	0.37	0.94	0.24	0.93	0.96	0.90	0.00
mCCD	1.00	0.38	1.00	0.95	0.98	1.00	0.78	1.00	1.00	1.00	0.30
mCC	1.00	1.00	0.82	1.00	1.00	1.00	1.00	1.00	1.00	0.81	0.65
mRCC1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
mRCC2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

perfect selection with the oracle variables, even though this is a nonlinear model and heavy-tailed data exists. The performance of DCS seems to fall behind.

Example 5. The following Poisson regression model is from Mai and Zou (2015), and we simply extend it to the multi-response case:

$$Y_j \sim \text{Poisson}(\mu_j), \quad \mu_j = \exp(0.8X_1 - 0.8X_2), \quad j = 1, \ldots, 10,$$

where the predictors $X_k \sim t_4$ independently, for $k = 1, 2, \ldots, 2000$. The sample size is 200 and $q = 10$.

The results are shown in Table 8. Surprisingly, the mRCC1 is still among the best, though the responses are discrete values with many ties and some extreme values. This implies that the mRCC1 may be suitable for regression problems with categorical data, while mRCC2 may not (the implementation of Kendall's τ correlation in mRRCS and mRCC2 uses formula (2.4) in Li et al. (2012), which may be inadequate for tied variables).

Table 8. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ and the proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ out of 200 replications in Example 5.

	$\mathcal{S}$					$\mathcal{P}_s$		$\mathcal{P}_a$
Method	5%	25%	50%	75%	95%	X1	X2	All
mSIS	2.0	3.0	15.5	136.0	1,346.5	0.85	0.86	0.70
mNIS	3.0	27.0	58.5	223.0	1,405.2	0.79	0.80	0.59
mRRCS	1,079.8	1,848.3	1,971.0	1,999.0	2,000.0	0.01	1.00	0.01
DCS	2.0	2.0	2.0	5.3	481.1	0.92	0.95	0.87
mCCD	2.0	2.0	2.0	2.0	2.0	1.00	1.00	1.00
mCC	2.0	6.0	16.0	47.5	664.9	0.88	0.92	0.80
mRCC1	2.0	2.0	2.0	2.0	2.0	1.00	1.00	1.00
mRCC2	1,208.4	1,882.5	1,980.0	1,999.3	2,000.0	0.01	1.00	0.01

5.2. Genomic data example

The breast cancer data set is described by Chin et al. (2006) and analyzed by Witten, Tibshirani and Hastie (2009), Chen, Dong and Chan (2013) and Molstad and Rothman (2016). The data set is publicly available in the R package PMA (Witten, Tibshirani and Hastie (2009)). It consists of gene expression measurements and comparative genomic hybridization measurements for $n = 89$ subjects. The goal is to explore the relationship between DNA copy-number variations and gene expression profiles, because certain types of cancer are characterized by unusual DNA copy-number changes, as shown in previous studies. Hence, we treat the DNA copy-number as the q-variate response, and the gene expression profile as the p-variate predictor. We conduct a multi-response regression analysis for chromosome 16, and its dimension is $(p, q) = (815, 61)$. Both the responses and predictors are standardized.

We include all of the aforementioned screening methods to carry out the multivariate response regression for the comparison. First, we randomly split 89 samples into training and test sets. Two proportions of training samples $\gamma = 0.5, 0.8$ are considered. Then, we apply each screening method to the training samples to select the top $d = 2[n_{\text{train}}/\log n_{\text{train}}]$ genes, where n_{train} is the training sample size. Moreover, we fit a multi-response Gaussian model using a "group-Lasso" penalty on the coefficients for each selected predictor after screening, and make predictions based on the test samples. The model fitting process and tuning parameter selection are implemented using the R package *glmnet*. Following Chen, Dong and Chan (2013), we calculate the mean squared prediction error $\|\mathbf{Y}_{\text{test}} - \mathbf{X}_{\text{test}}\widehat{\mathbf{B}}\|_{\text{F}}^2/(qn_{\text{test}})$, where $(\mathbf{Y}_{\text{test}}, \mathbf{X}_{\text{test}})$ denotes the test set, $\widehat{\mathbf{B}}$ denotes the estimated coefficient matrix, and n_{test} is the sample size of the test set. The above procedure is repeated 200 times.

Table 9. The means of the prediction errors in 200 times randomly split genomic data. The standard deviations of the prediction errors are shown in parentheses, where γ is the proportion of the training samples. NM is the null model using componentwise means of training responses to predict the test sample.

γ	mSIS	mNIS	mRRCS	DCS	mCCD	mCC	mRCC1	mRCC2	NM
0.5	0.769	0.758	0.791	0.765	0.781	0.786	0.774	0.738	1.021
	(0.066)	(0.068)	(0.065)	(0.068)	(0.067)	(0.08)	(0.079)	(0.067)	(0.089)
0.8	0.737	0.737	0.784	0.735	0.753	0.756	0.756	0.685	1.047
	(0.135)	(0.129)	(0.141)	(0.129)	(0.136)	(0.132)	(0.146)	(0.117)	(0.184)

Table 10. The p-values of two-sided paired samples t-test for the proposed methods against other methods.

γ		mSIS	mNIS	mRRCS	DCS	mCCD	mCC	mRCC1	mRCC2	NM
0.5	mCC	0.001	0	0.307	0	0.336	-	0.01	0	0
	mRCC1	0.286	0	0.001	0.072	0.201	0.01	-	0	0
	mRCC2	0	0	0	0	0	0	0	-	0
0.8	mCC	0.081	0.079	0.016	0.058	0.743	-	0.917	0	0
	mRCC1	0.094	0.092	0.015	0.072	0.785	0.917	-	0	0
	mRCC2	0	0	0	0	0	0	0	-	0

Table 11. Genes with top seven highest selection frequency by mRCC2.

$\gamma = 0.5$	genenames	COX4I1	KIAA0174	FLJ13868	KIAA1007	USP10	PARN	KATNB1
	frequency	0.82	0.76	0.72	0.58	0.57	0.56	0.55
$\gamma = 0.8$	genenames	COX4I1	FLJ13868	KIAA0174	SF3B3	KATNB1	KIAA1007	USP10
	frequency	1.00	1.00	0.99	0.97	0.97	0.96	0.95

The means of the prediction errors with their standard deviations are presented in Table 9. For each splitting ratio, mRCC2 enjoys outstanding predictive performance. To check whether the MSEs for the proposed approaches are significantly different from those for the other methods, we perform two-sided paired-sample t-tests for the mCC, mRCC1, and mRCC2 against other methods; the corresponding p-values are presented in Table 10. We also conduct a one-sided paired-sample t-test for mRCC2 only, and its p-values are still zero, which confirms that mRCC2 has significantly lower prediction errors than those of other methods. Furthermore, we list the genes with the top seven highest selection frequencies by mRCC2 in Table 11. We can see the top three among these are COX4I1, FLJ13868, and KIAA0174 for both splitting ratios. Therefore, in this example, mRCC2 may provide biological researchers with a more targeted list of gene expression profiles, which could be useful in subsequent studies.

Acknowledgments

We sincerely thank the editor, AE, and referees for their helpful comments. He was supported by the Youth Program of National Natural Science Foundation of China (11901289). Zhou's work was supported by the State Key Program of National Natural Science Foundation of China (71931004). Zou's work was supported, in part, by NSF DMS 1915842.

Appendix

A. Appendix

Lemma 1 (Theorem A, Serfling (1980), p. 201)**.** *Let $X_1, X_2, \ldots, X_n$ be independent observations on a distribution function F. Let $h = h(x_1, \ldots, x_m)$ be a kernel for a "parametric function" $\theta = \theta(F)$, with $a \leq h(x_1, \ldots, x_m) \leq b$. Put $\theta = E\{h(X_1, \ldots, X_m)\}$, then, for $t > 0$ and $n \geq m$,*

$$\Pr(U_n - \theta \geq t) \leq \exp\left(-\frac{2[n/m]t^2}{(b-a)^2}\right),$$

where U_n is the U-statistic corresponding to the kernel h for the estimation of θ, that is,

$$U_n = \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, \ldots, X_{i_m}) \tag{A.1}$$

with $\sum_c$ denotes summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $\{1, \ldots, n\}$.

Lemma 2. *Given a sample $(X_i, Y_i)_{i=1}^n$, for any $\delta > 0$, the Spearman correlation $\widehat{r}_s(X, Y)$ has the following tail bound*

$$\Pr\left(|\widehat{r}_s - r_s| \geq \frac{6}{n} + \delta\right) \leq 2\exp\left(-\frac{(n-3)(n+1)^2\delta^2}{24(n-2)^2}\right) \\ +2\exp\left(-\frac{(n-2)(n+1)^2\delta^2}{16}\right),$$

for $n > 3$, where r_s is the population Spearman correlation.

Proof of Lemma 2. If we take h in (A.1) to be the kernel of degree $m = 2$ given by

$$h_{\widehat{\tau}}\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right) = \operatorname{sgn}(x_1 - x_2)\operatorname{sgn}(y_1 - y_2),$$

then $\widehat{\tau} \triangleq U_{h_{\widehat{\tau}}}$ is the sample Kendall's tau correlation. If we take h in (A.1) to be the kernel of degree $m = 3$ given by

$$h_{\widetilde{r}_s}\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}\right) = \frac{1}{2} \sum_{\substack{i,j,l=1 \\ i\neq j, j\neq l, i\neq l}}^{3} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_l),$$

and define $\widetilde{r}_s \triangleq U_{h_{\widetilde{r}_s}}$,Hoeffding (1948) showed that

$$\widehat{r}_s = \frac{n-2}{n+1}\widetilde{r}_s + \frac{3}{n+1}\widehat{\tau}. \tag{A.2}$$

Hence, the dominating term $\widetilde{r}_s$ of Spearman correlation is a U-statistic. Since $\widetilde{r}_s, \widehat{\tau}$ are unbiased estimators of their population version r_s, τ respectively, and $|h_{\widetilde{r}_s}| \leq 1, |h_{\widehat{\tau}}| \leq 1$, by Lemma 1

$$\Pr\left(\widetilde{r}_s - r_s \geq \frac{\delta}{2}\right) \leq \exp\left(-\frac{(n-3)\delta^2}{24}\right), \tag{A.3}$$

$$\Pr\left(\widehat{\tau} - \tau \geq \frac{\delta}{2}\right) \leq \exp\left(-\frac{(n-2)\delta^2}{16}\right), \tag{A.4}$$

for any $\delta > 0$ and $n > 3$. Note that $-6 \leq 3(\tau - r_s) \leq 6$, we have

$$\begin{aligned}
&\Pr\left(|\widehat{r}_s - r_s| \geq \frac{6}{n} + \delta\right) \\
&\leq \Pr\left(\widehat{r}_s - r_s \geq \frac{6}{n+1} + \delta\right) + \Pr\left(\widehat{r}_s - r_s \leq -\frac{6}{n+1} - \delta\right) \\
&\leq \Pr\left(\widehat{r}_s - r_s - \frac{3(\tau - r_s)}{n+1} \geq \delta\right) + \Pr\left(\widehat{r}_s - r_s - \frac{3(\tau - r_s)}{n+1} \leq -\delta\right) \\
&\leq \Pr\left(\frac{n-2}{n+1}(\widetilde{r}_s - r_s) + \frac{3}{n+1}(\widehat{\tau} - \tau) \geq \delta\right) \\
&\quad + \Pr\left(\frac{n-2}{n+1}(\widetilde{r}_s - r_s) + \frac{3}{n+1}(\widehat{\tau} - \tau) \leq -\delta\right) \\
&\leq \Pr\left(\frac{n-2}{n+1}(\widetilde{r}_s - r_s) \geq \frac{\delta}{2}\right) + \Pr\left(\frac{3}{n+1}(\widehat{\tau} - \tau) \geq \frac{\delta}{2}\right) \\
&\quad + \Pr\left(\frac{n-2}{n+1}(\widetilde{r}_s - r_s) \leq -\frac{\delta}{2}\right) + \Pr\left(\frac{3}{n+1}(\widehat{\tau} - \tau) \leq -\frac{\delta}{2}\right) \\
&\leq 2\exp\left(- \frac{(n-3)(n+1)^2\delta^2}{24(n-2)^2}\right) + 2\exp\left(- \frac{(n-2)(n+1)^2\delta^2}{16}\right).
\end{aligned}$$

Throughout the rest of the paper, for any matrix $\mathbf{A}$, denote $\|\mathbf{A}\| = \sqrt{\lambda_{\max}(\mathbf{A}^{\mathrm{T}}\mathbf{A})}$ be the spectral norm and $\|\mathbf{A}\|_{\max} = \max_{i,j} |\mathbf{A}_{i,j}|$ be the max norm.

Lemma 3. *Under Condition* (C1)*, for any* $c_8 > 0$*, there exist some positive constants* c_3*,* c_4 *and* C*, such that for* $n > Cq^2$

$$\begin{aligned}&\Pr\left(\left|\|\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}^{-1}\| - \|\boldsymbol{\Sigma}_{R(\mathbf{Y})}^{-1}\|\right| \geq c_8\|\boldsymbol{\Sigma}_{R(\mathbf{Y})}^{-1}\|\right)\\ &\leq 2q^2\exp(-c_3nq^{-4}) + 2q^2\exp(-c_4n^3q^{-4}).\end{aligned}$$

Proof of Lemma 3. For any symmetric matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{D}$, by the similar argument in the proof of Lemma 5 of Fan, Feng and Song (2011), we have

$$\begin{aligned}|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \leq \max\{|\lambda_{\min}(\mathbf{A}-\mathbf{B})|, |\lambda_{\min}(\mathbf{B}-\mathbf{A})|\},\\ |\lambda_{\min}(\mathbf{D})| \leq d\|\mathbf{D}\|_{\max}, \quad |\lambda_{\min}(-\mathbf{D})| \leq d\|\mathbf{D}\|_{\max},\end{aligned}$$

where d is the dimension of $\mathbf{D}$. Hence,

$$|\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}) - \lambda_{\min}(\boldsymbol{\Sigma}_{R(\mathbf{Y})})| \leq q\|\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})} - \boldsymbol{\Sigma}_{R(\mathbf{Y})}\|_{\max}.$$

For any $\delta_1 > 0$, it follows from Lemma 2 that the union bound of probability

$$\begin{aligned}&\Pr\left(|\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}) - \lambda_{\min}(\boldsymbol{\Sigma}_{R(\mathbf{Y})})| \geq q\left(\frac{6}{n} + \delta_1\right)\right)\\ &\leq q^2\Pr\left(|\widehat{r}_s(Y_j, Y_l) - r_s(Y_j, Y_l)| \geq \frac{6}{n} + \delta_1\right)\\ &\leq 2q^2\exp(-\tilde{c}_3n\delta_1^2) + 2q^2\exp(-\tilde{c}_4n^3\delta_1^2),\end{aligned}$$

for some positive constant $\tilde{c}_3$ and $\tilde{c}_4$. Take $\delta_1 = c_9c_0q^{-2} - 6/n$ in the above, where $c_9 \in (0,1)$, denote $C = 6/(c_9c_0)$, by the Condition (C1), when $n > Cq^2$ it follows that

$$\begin{aligned}&\Pr\left(|\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}) - \lambda_{\min}(\boldsymbol{\Sigma}_{R(\mathbf{Y})})| \geq c_9\lambda_{\min}(\boldsymbol{\Sigma}_{R(\mathbf{Y})})\right)\\ &\leq 2q^2\exp(-c_3nq^{-4}) + 2q^2\exp(-c_4n^3q^{-4}),\end{aligned}$$

for some positive constant c_3 and c_4. If A and B are two positive constants, it is shown in the proof of Lemma 5 of Fan, Feng and Song (2011) that for $a \in (0,1)$,

$$|A^{-1} - B^{-1}| \geq cB^{-1} \quad \text{implies} \quad |A - B| \geq aB,$$

where $c = 1/(1-a) - 1$. Therefore, by the fact that $\lambda_{\min}^{-1}(\mathbf{D}) = \lambda_{\max}(\mathbf{D}^{-1})$, we have

$$
\begin{aligned}
&\Pr\left(\left|\|\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}^{-1}\| - \|\boldsymbol{\Sigma}_{R(\mathbf{Y})}^{-1}\|\right| \geq c_8\|\boldsymbol{\Sigma}_{R(\mathbf{Y})}^{-1}\|\right)\\
&= \Pr\left(|\lambda_{\min}^{-1}(\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}) - \lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{R(\mathbf{Y})})| \geq c_8\lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{R(\mathbf{Y})})\right)\\
&\leq \Pr\left(|\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}) - \lambda_{\min}(\boldsymbol{\Sigma}_{R(\mathbf{Y})})| \geq c_9\lambda_{\min}(\boldsymbol{\Sigma}_{R(\mathbf{Y})})\right)\\
&\leq 2q^2\exp(-c_3nq^{-4}) + 2q^2\exp(-c_4n^3q^{-4}),
\end{aligned}
$$

where $c_8 = 1/(1-c_9) - 1 > 0$.

Proof of Theorem 1. We only focus on the proof for mRCC1, since the proof for mRCC2 is similar by modifying tail probability using (A.4) in Lemma 3 and the following. For the first part of the theorem, recall that

$$
\widehat{\omega}_k^{\text{mRCC1}} = \widehat{\mathbf{r}}_k^{\text{T}}\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}^{-1}\widehat{\mathbf{r}}_k,
$$

$$
\omega_k^{\text{mRCC1}} = \mathbf{r}_k^{\text{T}}\boldsymbol{\Sigma}_{R(\mathbf{Y})}^{-1}\mathbf{r}_k,
$$

we have

$$
\begin{aligned}
\widehat{\omega}_k^{\text{mRCC1}} - \omega_k^{\text{mRCC1}} &= (\widehat{\mathbf{r}}_k - \mathbf{r}_k)^{\text{T}}\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}^{-1}(\widehat{\mathbf{r}}_k - \mathbf{r}_k)\\
&\quad +2(\widehat{\mathbf{r}}_k - \mathbf{r}_k)^{\text{T}}\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}^{-1}\mathbf{r}_k\\
&\quad +\mathbf{r}_k^{\text{T}}(\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}^{-1} - \boldsymbol{\Sigma}_{R(\mathbf{Y})}^{-1})\mathbf{r}_k\\
&\triangleq I_1 + I_2 + I_3.
\end{aligned}
$$

Note that

$$
I_1 \leq \|\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}^{-1}\| \cdot \|\widehat{\mathbf{r}}_k - \mathbf{r}_k\|^2.
$$

By Lemma 2, for any $\delta > 0$, the union bound of probability is

$$
\begin{aligned}
\Pr\left(\|\widehat{\mathbf{r}}_k - \mathbf{r}_k\|^2 \geq q\left(\frac{6}{n} + \delta\right)^2\right) &\leq q\Pr\left(|\widehat{r}_s(X_k, Y_j) - r_s(X_k, Y_j)|^2 > \left(\frac{6}{n} + \delta\right)^2\right)\\
&\leq 2q\exp(-\tilde{c}_5n\delta^2) + 2q\exp(-\tilde{c}_6n^3\delta^2),
\end{aligned}
$$

for some positive constant $\tilde{c}_5$ and $\tilde{c}_6$. Under Condition (C1),

$$
\|\boldsymbol{\Sigma}_{R(\mathbf{Y})}^{-1}\| \leq c_0^{-1}q.
$$

By Lemma 3,

$$
\Pr\left(\|\widehat{\boldsymbol{\Sigma}}_{R(\mathbf{Y})}^{-1}\| \geq (c_8+1)c_0^{-1}q\right) \leq 2q^2\exp(-c_3nq^{-4}) + 2q^2\exp(-c_4n^3q^{-4}),
$$

for any $c_8 > 0$, $n > Cq^2$ and some positive constants c_3, c_4 and C. Hence, the union bound of probability for I_1 is

$$\begin{aligned}\Pr\left(|I_1| \geq (c_8+1)c_0^{-1}q^2\left(\frac{6}{n}+\delta\right)^2\right) \leq\ & 2q^2\exp(-c_3nq^{-4}) + 2q^2\exp(-c_4n^3q^{-4})\\ & +2q\exp(-\tilde{c}_5n\delta^2) + 2q\exp(-\tilde{c}_6n^3\delta^2).\end{aligned}$$

We next deal with the probability bound for I_2. Note that

$$|I_2| \leq 2\|(\widehat{\mathbf{r}}_k - \mathbf{r}_k)^{\mathrm{T}}\| \cdot \|\widehat{\mathbf{\Sigma}}_{R(\mathbf{Y})}^{-1}\| \cdot \|\mathbf{r}_k\|.$$

It is obvious that

$$\|\mathbf{r}_k\|^2 \leq q.$$

Hence, the union bound of probability for I_2 is

$$\begin{aligned}\Pr\left(|I_2| \geq 2(c_8+1)c_0^{-1}q^2\left(\frac{6}{n}+\delta\right)\right) \leq\ & 2q^2\exp(-c_3nq^{-4}) + 2q^2\exp(-c_4n^3q^{-4})\\ & +2q\exp(-\tilde{c}_5n\delta^2) + 2q\exp(-\tilde{c}_6n^3\delta^2).\end{aligned}$$

To bound I_3, note that

$$I_3 = \mathbf{r}_k^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{R(\mathbf{Y})}^{-1}(\mathbf{\Sigma}_{R(\mathbf{Y})} - \widehat{\mathbf{\Sigma}}_{R(\mathbf{Y})})\mathbf{\Sigma}_{R(\mathbf{Y})}^{-1}\mathbf{r}_k.$$

By the fact that $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$, we have

$$|I_3| \leq \|\widehat{\mathbf{\Sigma}}_{R(\mathbf{Y})}^{-1}\| \cdot \|\mathbf{\Sigma}_{R(\mathbf{Y})} - \widehat{\mathbf{\Sigma}}_{R(\mathbf{Y})}\| \cdot \|\mathbf{\Sigma}_{R(\mathbf{Y})}^{-1}\| \cdot \|\mathbf{r}_k\|^2.$$

For a d-dimensional square matrix $\mathbf{D}$, it is shown in the proof of Lemma 5 of Fan, Feng and Song (2011) that $\|\mathbf{D}\| \leq d\|\mathbf{D}\|_{\max}$. Therefore,

$$\begin{aligned}&\Pr\left(\|\mathbf{\Sigma}_{R(\mathbf{Y})} - \widehat{\mathbf{\Sigma}}_{R(\mathbf{Y})}\| \geq q\left(\frac{6}{n}+\delta\right)\right)\\ &\leq q^2\Pr\left(|r_s(Y_j,Y_l) - \widehat{r}_s(Y_j,Y_l)| \geq \frac{6}{n}+\delta\right)\\ &\leq 2q^2\exp(-\tilde{c}_5n\delta^2) + 2q^2\exp(-\tilde{c}_6n^3\delta^2).\end{aligned}$$

Hence, the union bound of probability for I_3 is

$$\begin{aligned}\Pr\left(|I_3| \geq (c_8+1)c_0^{-2}q^4\left(\frac{6}{n}+\delta\right)\right) \leq & 2q^2\exp(-c_3nq^{-4}) + 2q^2\exp(-c_4n^3q^{-4})\\ & + 2q^2\exp(-\tilde{c}_5n\delta^2) + 2q^2\exp(-\tilde{c}_6n^3\delta^2).\end{aligned}$$

The final probability bound

$$\Pr\left(|\widehat{\omega}_k^{\mathrm{mRCC1}} - \omega_k^{\mathrm{mRCC1}}| \geq c_{10}q^2\left(\frac{6}{n}+\delta\right)^2 + 2c_{10}q^2\left(\frac{6}{n}+\delta\right) + c_{11}q^4\left(\frac{6}{n}+\delta\right)\right)$$

$$\leq 6q^2\Big(\exp(-c_3nq^{-4})+\exp(-c_4n^3q^{-4})\Big) \\ +(2q^2+4q)\Big(\exp(-\tilde{c}_5n\delta^2)+\exp(-\tilde{c}_6n^3\delta^2)\Big),$$

for some positive constants c_{10} and c_{11}. Take $\delta = n^{-\kappa}-6/n$, when $n > \max\{Cq^2, 6^{1/(1-\kappa)}\}$, there exists $c_2 > 0$, such that $c_{10}q^2n^{-2\kappa}+2c_{10}q^2n^{-\kappa}+c_{11}q^4n^{-\kappa}=c_2q^4n^{-\kappa}$ and

$$\Pr\left(|\widehat{\omega}_k^{\mathrm{mRCC1}}-\omega_k^{\mathrm{mRCC1}}|\geq c_2q^4n^{-\kappa}\right) \\ \leq 6q^2\Big(\exp(-c_3nq^{-4})+\exp(-c_4n^3q^{-4})\Big) \\ +(2q^2+4q)\Big(\exp(-c_5n^{1-2\kappa})+\exp(-c_6n^{3-2\kappa})\Big),$$

for some positive constants c_5 and c_6. Thus the first part immediately follows the union bound of probability.

Next, we show the second part of the theorem. By Condition (C2), under the event

$$\Gamma_n=\left\{\max_{k\in j=1,\ldots,p}|\widehat{\omega}_k^{\mathrm{mRCC1}}-\omega_k^{\mathrm{mRCC1}}|\leq\frac{\delta_{\widetilde{\mathcal{A}}}q^4n^{-\kappa}}{2}\right\},$$

we have

$$\min_{k\in\widetilde{\mathcal{A}}}\widehat{\omega}_k^{\mathrm{mRCC1}}\geq\min_{k\in\widetilde{\mathcal{A}}}\{\omega_k^{\mathrm{mRCC1}}-|\widehat{\omega}_k^{\mathrm{mRCC1}}-\omega_k^{\mathrm{mRCC1}}|\} \\ \geq\max_{k\in\widetilde{\mathcal{A}}^c}\omega_k^{\mathrm{mRCC1}}+\frac{\delta_{\widetilde{\mathcal{A}}}q^4n^{-\kappa}}{2}\geq\max_{k\in\widetilde{\mathcal{A}}^c}\widehat{\omega}_k^{\mathrm{mRCC1}}$$

Hence, there must exists $\nu_n\geq t_n$, such that $\widetilde{\mathcal{A}}=\widehat{\mathcal{A}}_{\nu_n}$. Moreover, for any $t_n\leq\nu_n$, $\widehat{\mathcal{A}}_{\nu_n}\subset\widehat{\mathcal{A}}_{t_n}$, which implies $\mathcal{A}\subset\widetilde{\mathcal{A}}\subset\widehat{\mathcal{A}}_{t_n}$. Therefore, let $c_2=\delta_{\mathcal{A}}/2$, by the choice of $t_n=c_1q^4n^{-\kappa}, c_1\leq\delta_{\mathcal{A}}/2$, we have

$$P(\mathcal{A}\subset\widehat{\mathcal{A}}_{t_n})\geq P(\Gamma_n)\geq 1-p\cdot\Big\{6q^2\Big(\exp(-c_3nq^{-4})+\exp(-c_4n^3q^{-4})\Big) \\ +(2q^2+4q)\Big(\exp(-c_5n^{1-2\kappa})+\exp(-c_6n^{3-2\kappa})\Big)\Big\}.$$

Proof of Theorem 2. We only focus on the proof for mRCC1, since the proof for mRCC2 is similar. Under Condition (C1),

$$\sum_{k=1}^p\omega_k^{\mathrm{mRCC1}}\leq\sum_{k=1}^p\|\mathbf{\Sigma}_{R(\mathbf{Y})}^{-1}\|\cdot\|\mathbf{r}_k\|^2\leq c_0^{-1}sq^2=O(sq^2).$$

This indicates that the number of $\{k:\omega_k^{\mathrm{mRCC1}}>\epsilon q^4n^{-\kappa}\}$ cannot exceed $O(sq^{-2}n^{\kappa})$

for any $\epsilon > 0$. Therefore, on the set

$$\Delta_n = \Big\{ \max_{1 \le k \le p} |\widehat{\omega}_k^{\mathrm{mRCC1}} - \omega_k^{\mathrm{mRCC1}}| \le \epsilon q^4 n^{-\kappa} \Big\},$$

the number of $\{k : \widehat{\omega}_k^{\mathrm{mRCC1}} > 2\epsilon q^4 n^{-\kappa}\}$ cannot exceed the number of $\{k : \omega_k^{\mathrm{mRCC1}} > \epsilon q^4 n^{-\kappa}\}$, which is bounded by $O(sq^{-2}n^{\kappa})$. Take $\epsilon = c_1/2$, we have

$$\Pr\big(|\widehat{\mathcal{A}}_{t_n}| \le O(sq^{-2}n^{\kappa})\big) \ge \Pr(\Delta_n).$$

The conclusion follows from the first part of Theorem 1.

References

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* **22**, 327–351.

Bergamaschi, A., Kim, Y. H., Kwei, K. A., Choi, Y., Bocanegra, M., Langerod, A. et al. (2008). CAMK1D amplification implicated in epithelial–mesenchymal transition in basal-like breast cancer. *Molecular Oncology* **2**, 327–339.

Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics* **41**, 2123–2148.

Chang, J., Tang, C. Y. and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *The Annals of Statistics* **44**, 515–539.

Chen, K., Dong, H. and Chan, K. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika* **100**, 901–920.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association* **107**, 1533–1545.

Chin, K., Devries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L. et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541.

Cui, H., Li, R. and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association* **110**, 630–641.

Durbin, J. and Stuart, A. (1951). Inversions and rank correlation coefficients. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **13**, 303–309.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics* **36**, 2605–2637.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106**, 544–557.

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

Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association* **109**, 1270–1284.

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

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics* **41**, 342–369.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**, 293–325.

Huang, D., Li, R. and Wang, H. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business & Economic Statistics* **32**, 237–244.

Kong, E., Xia, Y. and Zhong, W. (2019). Composite coefficient of determination and its application in ultrahigh dimensional variable screening. *Journal of the American Statistical Association* **114**, 1740–1751.

Lee, W. and Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Journal of Multivariate Analysis* **111**, 241–255.

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

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

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association* **109**, 266–274.

Mai, Q. and Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100**, 229–234.

Mai, Q. and Zou, H. (2015). The fused Kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics* **43**, 1471–1497.

Molstad, A. J. and Rothman, A. J. (2016). Indirect multivariate response linear regression. *Biometrika* **103**, 595–607.

Obozinski, G., Taskar, B. and Jordan, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing* **20**, 231–252.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R. et al. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics* **4**, 53–77.

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E. et al. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12963–12968.

Rothman, A. J., Levina, E. and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19**, 947–962.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10869–10874.

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

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

Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. D. C. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of The Royal Statistical Society. Series B (Statistical Methodology)* **69**, 329–346.

Zhao, H., Langerod, A., Ji, Y., Nowels, K. W., Nesland, J. M., Tibshirani, R. et al. (2004). Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Molecular Biology of the Cell* **15**, 2523–2536.

Zhu, L., Li, L., Li, R. and Zhu, L. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

Di He

School of Economics, Nanjing University, Nanjing, 210046, China.

E-mail: hedi@nju.edu.cn

Yong Zhou

Key Laboratory of Advanced Theory and Application in Statistics and Data Science, MOE, and Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai 200062, China.

E-mail: yzhou@amss.ac.cn

Hui Zou

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: zouxx019@umn.edu

(Received November 2018; accepted December 2019)