On Sure Screening with Multiple Responses

Di He$^1$, Yong Zhou$^2$ and Hui Zou$^3$

$^1$ Nanjing University, $^2$ East China Normal University and $^3$ 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 sure independence screening paper by Fan and Lv (2008), many variable screening methods have been proposed and studied in the literature. Yet, the majority of the existing screening methods handle the classical univariate response data case and do not apply naturally to the multiple responses datasets. In this paper, 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 named multi-response rank canonical correlation screening (mRCC). It 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 regular-
ity conditions. Extensive numerical experiments also demonstrate the superior performances of mRCC over other available alternatives.

Key words and phrases: Multi-response data; Rank correlation; Canonical correlation; Sure screening property.

1. Introduction

Multivariate responses are commonly encountered in many statistical applications. For example, microarray expression experiments and array CGH (comparative genomic hybridization) experiments have been conducted by biologists in breast cancer cohort studies [Sorlie 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, respectively. Although the analysis of expression arrays alone or CGH arrays alone has provided useful information, an integrative analysis of DNA copy numbers and the gene expression files are necessary as these two types of data offer complementary information. Hence, integrating DNA and RNA data benefits the recognition of more subtle genetic regulatory relationships in cancer cells [Pollack et al., 2002].

A straightforward way to model and analyze such datasets is the multiresponse 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

\[
Y = B_0 + XB + E, 
\]

where \( Y = (Y_1, \cdots, Y_q) \) is an \( n \times q \) response matrix, \( X = (X_1, \cdots, X_p) \) is an \( n \times p \) design matrix, \( B = (\beta_{kj}) \) is an \( p \times q \) matrix of parameters, \( B_0 = (\beta_{01}, \cdots, \beta_{0q}) \) is an \( n \times q \) matrix of intercepts with \( 1 \) a \( n \)-vector whose entries all being 1, and \( E \) is an unobserved \( n \times q \) matrix whose row vectors \( \epsilon_1, \cdots, \epsilon_n \) are independent copies with mean zero and covariance matrix \( \Sigma_E \). In general we should not treat the multi-response problem as multiple univariate response problems, although sometimes the solutions may be the same. For example, we can obtain the ordinary least squares estimator of (1.1) by performing separate linear regression on each response. If the errors are correlated, a weighted criterion of residual sum of squares arises naturally and the solution still amounts to the ordinary least squares estimates. However, this is not the case for regression with LASSO penalty on the entries of \( B \). When LASSO regression involves a known \( \Sigma_E \), the optimal solution for \( B \) obtained from the weighted criterion accounts for the inverse of \( \Sigma_E \) (Rothman et al., 2010), which is different from the separate lasso regression estimates with each response. When \( p \) is very large, there are challenges of computational efficiency, statistical consistency, and algorithmic stability (Fan et al., 2009). To this end, many shrinkage estimators
of the parameters have been proposed for multi-response regression (1.1), by penalizing the optimization with the residual sum of squares. Some simultaneously estimate the parameters and discard irrelevant predictors using proper regularization (Obozinski et al., 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) forcefully argued that it is beneficial for both computations and theoretical considerations to first reduce the ambient dimension to a moderately high dimension and then fit a regularized model. The dimension reduction step should preserve all important features—a property known as sure screening property. To demonstrate their philosophy, Fan and Lv (2008) introduced a sure independence screening (SIS) procedure by using 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 have proposed more sophisticated screening methods to deal with more complicated models, such as, maximum marginal likelihood screening for generalized linear models (Fan et al., 2010), the nonparametric independence screening (NIS) for additive models (Fan et al., 2011), the robust
rank correlation screening (RRCS) for semiparametric single-index models with a monotonic link function \( \text{Li et al., 2012a} \), the quantile-adaptive screening for quantile regression \( \text{He et al., 2013} \), the empirical likelihood screening for the parametric models that can be formulated via general estimating equations \( \text{Chang et al., 2013} \), and so on. \( \text{Fan et al., 2014} \) extended NIS for varying coefficient models while \( \text{Liu et al., 2014} \) considered this type of models based on conditional correlation coefficient. \( \text{Chang et al., 2016} \) proposed a unified approach for nonparametric and semiparametric models by marginal empirical likelihood. When the response is binary, \( t \)-statistic was proposed by \( \text{Fan and Fan, 2008} \) to screen predictors, while \( \text{Mai and Zou, 2013} \) developed the Kolmogorov filter using Kolmogorov-Smirnov statistic. \( \text{Huang et al., 2014} \) proposed a Pearson chi-square based feature screening method for categorical response and predictors. \( \text{Cui et al., 2015} \) further considered discriminant analysis with multi-categorical response variable. Another popular screening genre is the model-free methods which overcome the model misspecification problem. For instance, the sure independent ranking and screening (SIRS) \( \text{Zhu et al., 2011} \), the distance correlation screening (DCS) \( \text{Li et al., 2012b} \) and the fused Kolmogorov filter \( \text{Mai and Zou, 2015} \).

The focus of this paper is multi-response data. The aforementioned
screening methods are primarily developed for the univariate data case. To our best knowledge, the only method that can naturally handle multivariate responses and univariate response is the distance correlation screening (DCS) because 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 paper is to develop new variable screening methods for multi-response data. First, we extend several existing screening methods (SIS, NIS, RRCS) into 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 problem. We believe a better variable screening method is possible if we take into account potential dependence among 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 correlation and/or Kendall’s $\tau$ correlation; with canonical correlation. It not only inherits multivariate merits of canonical correlation that takes advantage of the dependence structure among the multivariate responses, but also preserves nice properties of the rank correlation that can handle heavy-tailed predictors and responses as well as invariance against monotonic transformation of them. 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, mRCC is the recommended method in this paper for variable screening with multi-response data.

The rest of the paper is organized as follows. The extensions of several existing screening methods are given in Section 2. In Section 3 we first introduce a screening method based on canonical correlation and then propose mRCC. Theories established in Section 4 show that the sure screening property of mRCC holds under weak regularity conditions. Section 5 displays simulation experiments and a genomic data example for the comparison of all the methods. Technical proofs are presented in the Appendix.

**Notation.** Throughout the paper, we assume $X$ is centered to have mean 0 columnwise. We denote $X_k, Y_j$ as the $k$-th, $j$-th column of $X, Y$
for $k = 1, \cdots, p$ and $j = 1, \cdots, q$. To avoid introducing more 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 $Y$ to denote the random vector of the response. Let $\| \cdot \|$ be the Euclidean norm for a vector, and let $\| \cdot \|_F$ be the Frobenius norm for a matrix. Denote $\text{RSS}$ 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 sure independence screening (SIS) by using the marginal correlation ranking $X_k^T Y / \|X_k\| \|Y\|$ to filter out the features that are weakly correlated with the response. SIS can be viewed from a marginal regression perspective:

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

Under the condition that $X_k$’s are further standardized to have norm one, it is easy to show that it 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 $\text{RSS}$ of the marginal regression. To carry out a similar screening procedure when the response is multivariate, a straightforward idea is to generalize (2.1) into

$$\min_{B_0, \beta_k} \|Y - B_0 - X_k \beta_k\|_F^2. \tag{2.2}$$
where $\beta_k = (\beta_{k1}, \cdots, \beta_{kq})$ is a row vector of parameters. The $RSS$ has the following form:

$$RSS_k = \sum_{j=1}^{q} \|Y_{jc}\|^2 \cdot (1 - \hat{\rho}_{kj}^2),$$

where $Y_{jc} = Y_j - \bar{Y}_j$ and $\hat{\rho}_{kj} = \hat{\rho}(X_k, Y_j)$ is the sample Pearson correlation coefficient between $X_k$ and $Y_j$. We scale $Y$ and $X$ to have mean zero and norm one columnwise, in order to remove the scale influence. In this case, ranking according to the following three quantities is still equivalent: the $\ell_2$ norm of the coefficient vector, the sum of squares of Pearson correlation coefficients, or the descending order of the $RSS$. Hence, by aggregating the squares of Pearson correlation coefficients of the predictor with each response, we obtain

$$\hat{\omega}_k^{mSIS} = \|\hat{\rho}_k\|^2,$$

(2.3)

where $\hat{\rho}_k = (\hat{\rho}_{k1}, \cdots, \hat{\rho}_{kq})^T$, and refer to (2.3) as the multi-response sure independence screening (mSIS) statistic. Note that this approach is actually treating multi-response problem as multiple univariate response data problem. 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 et al. (2011) developed nonparametric independence screening (NIS)
for additive models that allows the true regression function to be nonlinear in the predictors. They considered the marginal nonparametric regression using basis function expansion such as B-splines. Similar as the generalization of SIS, we aggregate $RSS_k$ from the marginal nonparametric regressions with each response

$$\min_{f_k \in \mathcal{S}_n} \| Y - f_k(X_k) \|_F^2 = \min_{b_k \in \mathbb{R}^{d_n \times q}} \| Y - \Psi_k b_k \|_F^2, \quad (2.4)$$

where $f_k = (f_{k1}, \cdots, f_{kq})$ and $f_{kj}(X_k) = \sum_{l=1}^{d_n} \gamma_{kl} \psi_l(X_k)$ is a $n$-vector sample version intending to approximate $E(Y_j | X_k)$, $\mathcal{S}_n$ is the space of polynomial splines, $\Psi_k \triangleq (\psi_1(X_k), \cdots, \psi_{d_n}(X_k))$ denotes a $n \times d_n$ normalized B-spline basis matrix, $b_k = (\gamma_{k1}, \cdots, \gamma_{kq})$ and $\gamma_{kj} = (\gamma_{kj1}, \cdots, \gamma_{kJd_n})^T$, $j = 1, \cdots, q$. The corresponding solution is

$$\hat{f}_k(X_k) = \Psi_k (\Psi_k^T \Psi_k)^{-1} \Psi_k^T Y.$$

We can treat

$$\hat{\omega}_m^{\text{mNIS}} = \| \hat{f}_k(X_k) \|_F^2 \quad (2.5)$$

as the marginal utility of the multi-response nonparametric independence screening (mNIS) for $X_k$. Or equivalently, we can rank the predictors by the 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. (2012a) proposed to use Kendall’s $\tau$ correlation coefficient as a ranking statistic. Their method is named RRCS. The marginal utility they proposed equals a quarter of Kendall’s $\tau$ correlation coefficient, that is

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

Similar to (2.3), we try to extend it into multiple responses case by simply summing up the squares of Kendall’s $\tau$ correlations between the predictor and each response

$$\hat{\omega}_{mRRCS}^2 = \|\hat{\tau}_k\|^2, \quad (2.6)$$

where $\hat{\tau}_k = (\hat{\tau}(X_k, Y_1), \cdots, \hat{\tau}(X_k, Y_q))^T$. We refer to (2.6) as the multivariate robust rank correlation screening (mRRCS) statistic. The population version of Kendall’s $\tau$ correlation will be 0 if two random variables are independent, as a result $\omega_{mRRCS}^2$ will be 0 if $X_k$ is independent of the multivariate responses.

2.4 Distance correlation screening

Distance correlation (DC) (Székely et al., 2007) is capable of measuring dependence between two random vectors. Unlike the Person correlation and the Kendall’s $\tau$ correlation, distance correlation equals zero if and only if two random vectors are independent. This unique property motivated Li...
et al. (2012b) to consider distance correlations screening (DCS) and DCS is one of the most popular model-free variable screening methods. Distance correlation and DCS can be naturally applied to multi-response data. For sake of completeness, we briefly review distance correlation screening here.

The distance correlation can be computed through distance covariance. For a given sample \( \{ U_i, V_i \}_{i=1}^n \) from two random vectors \( U, V \), the squared distance covariance can be estimated as

\[
\hat{d}\text{cov}^2(U, V) = \hat{S}_1(U, V) + \hat{S}_2(U, V) - 2\hat{S}_3(U, V),
\]

where

\[
\hat{S}_1(U, V) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \| U_i - U_j \| \| V_i - V_j \|,
\]

\[
\hat{S}_2(U, V) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \| U_i - U_j \| \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \| V_i - V_j \|,
\]

\[
\hat{S}_3(U, V) = \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \| U_i - U_k \| \| V_j - V_k \|.
\]

Therefore, the distance correlation screening (DCS) can be implemented by computing

\[
\hat{\omega}^{\text{DCS}}_k = \hat{d}\text{cor}(X_k, Y) = \hat{d}\text{cov}(X_k, Y) / \sqrt{\hat{d}\text{cov}(X_k, X_k)\hat{d}\text{cov}(Y, Y)}.
\]

for each predictor \( X_k \).

Numerical studies (Li et al., 2012b) showed DCS can have good performances for very complex models. In general, DCS outperforms SIS unless
the true model is a linear regression model. Li et al. (2012b) proved that the sure screening property of the DCS holds if the responses and predictors are sub-Gaussian. On the other hand, in the presence of heavy-tailed data, numerical performances of DCS can be very poor (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 DCS is not invariant against monotonic transformation, while some variable screening methods are invariant under monotonic transformation, including the rank correlation screening (Li et al. 2012a) 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 canonical correlation analysis (CCA) and then introduce a novel way to combine rank correlation and canonical correlation.

3.1 Canonical correlation

Canonical correlation analysis is a way of inferring information from cross-covariance matrices which finds two projections for two random vectors such that the projected random vectors have maximum correlation with each other. For the $k$-th predictor $X_k$, the canonical correlation between
X_k and (Y_1, \cdots, Y_q) is defined as

\[
\rho^c_k = \max_b \frac{\Sigma^*_X Y b}{\sqrt{\Sigma^*_{Xk} \sqrt{b^T \Sigma^*_Y b}}},
\]

where Σ^*_X, Σ^*_X Y, Σ^*_Y are the submatrices of Σ^* =

\[
\begin{pmatrix}
Σ^*_X & Σ^*_{XkY} \\
Σ^*_{Xk} & Σ^*_Y
\end{pmatrix},
\]

which is the covariance matrix of \((X_k, (Y_1, \cdots, Y_q))\). By the Cauchy-Schwartz inequality, it can be shown that

\[
\rho^c_k = \left(\Sigma^*_X Y \Sigma^*_X Y^{-1} \Sigma^*_{XkY} \right)^{1/2} \left(\Sigma^*_{Xk} \right)^{1/2}
\]

(3.1)

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

\[
(\rho^c_k)^2 = \rho^T_k (\Sigma_Y)^{-1} \rho_k,
\]

(3.2)

recalling that \(\rho_k = (\rho_{k1}, \cdots, \rho_{kq})^T\) are the Pearson correlations between \(X_k\) and \(Y_j\)'s, and \(\Sigma_Y = (\rho_{jl})_{q \times q}\) is the correlation matrix of \((Y_1, \cdots, Y_q)\). We can use

\[
\hat{\omega}^{mCC}_k = (\hat{\rho}^c_k)^2 = \hat{\rho}^T_k (\hat{\Sigma}_Y)^{-1} \hat{\rho}_k
\]

(3.3)

as a variable screening statistic, which is called the multi-response canonical correlation screening (mCC) statistic in this work.

To compare mCC and mSIS, we see that \(\omega^{mSIS}_k\) simply sums up \(\rho^2_{kj}\) while (3.2) is a weighted summation of \(\rho^2_{kj}\) and \(\rho_{kj}\rho_{kl}, (j \neq l)\), recruiting
the information of cross-correlation among $Y_j$'s. 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 an generalization that can capture 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 attempt to consider a better version of mCC that integrates rank correlation with canonical correlation. Two commonly used rank correlations, Kendall’s $\tau$ correlation and Spearman rank correlation, are employed.

We first consider to replace the Pearson correlations between $X_k$ and $Y_j$’s in (3.3) with corresponding Spearman rank correlations

$$
(\hat{r}_k^c)^2 \triangleq \hat{r}_k^T \hat{\Sigma}^{-1}_R(Y) \hat{r}_k,
$$

where $\hat{r}_k = (\hat{r}_s(X_k, Y_1), \cdots, \hat{r}_s(X_k, Y_q))^T$ are the Spearman correlation coefficients between $X_k$ and $Y_j$’s, and $\hat{\Sigma}_R(Y) = (\hat{r}_s(Y_j, Y_l))_{q \times q}$ is a matrix of Spearman correlations among all pairs of $Y_j$ and $Y_l$. Since the Spearman rank correlation \cite{Spearman1904, DurbinStuart1951} that measures an ordinal association is analogous to the Pearson correlation between the rank values of two variables, $\hat{\Sigma}_R(Y)$ is exactly the sample correlation matrix
of \((R(Y_1), \cdots, R(Y_q))\), where \(R(\cdot)\) stands for the rank of a random variable among \(n\) observations.

Next, we adopt Kendall’s \(\tau\) correlation in (3.4) and define

\[
(\hat{\tau}_k^m)^2 \triangleq \hat{\tau}_k^T \hat{\Sigma}_{R(Y)}^{-1} \hat{\tau}_k, \tag{3.5}
\]

recalling that \(\hat{\tau}_k = (\hat{\tau}(X_k, Y_1), \cdots, \hat{\tau}(X_k, Y_q))^T\) are the Kendall’s \(\tau\) correlations between \(X_k\) and \(Y_j\)'s, and \(\hat{\Sigma}_{R(Y)} = (\hat{\tau}(Y_j, Y_l))_{q \times q}\) is a matrix of Kendall’s \(\tau\) correlations among all pairs of \(Y_j\) and \(Y_l\).

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

\[
\hat{A}_{t_n} = \{1 \leq k \leq p : \hat{\omega}_m^{mRCC1} \geq t_n\}, \quad \text{or} \quad \hat{A}_{t'_n} = \{1 \leq k \leq p : \hat{\omega}_m^{mRCC2} \geq t'_n\}
\]

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

**Remark 1.** The proposed mRCC not only inherits multivariate merits of canonical correlation that takes 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 transformation of them. Because of these nice properties and its excellent numerical performance in Section 5, mRCC is the main method we advocate to use in practice.

4. The Sure Screening Property

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

\[ A = \{ k : F(Y | X_k) functionally depends on X_k for some Y \}, \]

with size \( s = |A| \). For variable \( X_k \), the population versions of mRCC1 and mRCC2 are

\[ \omega_{mRCC1}^k = (r_k^c)^2 = r_k^T \Sigma_{R(Y)}^{-1} r_k, \]

(4.1)

and

\[ \omega_{mRCC2}^k = (\tau_k^c)^2 = \tau_k^T \Sigma_{\tilde{R}(Y)}^{-1} \tau_k, \]

(4.2)

respectively, where \( r_k = (r_s(X_k, Y_1), \ldots, r_s(X_k, Y_q))^T \), \( \Sigma_{R(Y)} = (r_s(Y_j, Y_l))_{q \times q} \) and \( \tau_k = (\tau(X_k, Y_1), \ldots, \tau(X_k, Y_q))^T \), \( \Sigma_{\tilde{R}(Y)} = (\tau(Y_j, Y_l))_{q \times q} \). For two random variables \( U \) and \( V \) from a joint distribution, let \( (U_1, V_1), (U_2, V_2), (U_3, V_3) \) be three independent realizations, then \( r_s(U, V) = \text{cov}(\text{sgn}(U_1 - U_2), \text{sgn}(V_1 - V_3)) \) and \( \tau(U, V) = \text{cov}(\text{sgn}(U_1 - U_2), \text{sgn}(V_1 - V_2)). \)

We consider the following conditions:
(C1) There exists a positive $c_0$ such that $\lambda_{\min}(\Sigma R(Y)) \geq c_0 q^{-1}$;

(C2) There exists $\tilde{A}$, a subset of \{1, \ldots, p\}, a constant $0 < \kappa < \frac{1}{2}$, such that

$$|\tilde{A}| \leq |\tilde{A}_n|, A \subset \tilde{A} \text{ and } \delta_{\tilde{A}} = q^{-4} n^\kappa \{\min_{k \in \tilde{A}} \omega_k^{mRCC1} - \max_{k \in \tilde{A}^c} \omega_k^{mRCC1}\} > 0;$$

(C1') There exists a positive $c'_0$ such that $\lambda_{\min}(\Sigma_{\tilde{R}}(Y)) \geq c'_0 q^{-1}$;

(C2') There exists $\tilde{A}'$, a subset of \{1, \ldots, p\}, a constant $0 < \kappa < \frac{1}{2}$, such that

$$|\tilde{A}'| \leq |\tilde{A}'_n|, A \subset \tilde{A}' \text{ and } \delta_{\tilde{A}'} = q^{-4} n^\kappa \{\min_{k \in \tilde{A}'} \omega_k^{mRCC2} - \max_{k \in \tilde{A}'^c} \omega_k^{mRCC2}\} > 0$$

Condition (C1) or (C1') rules out the case that one component of the multivariate responses is a perfect monotonic function of another with a perfect Spearman correlation of +1 or −1, or, the agreement or the disagreement between two rankings of two components of the response is perfect with a perfect Kendall’s $\tau$ correlation of +1 or −1, respectively. Condition (C2) or (C2') is very common in screening literature (Mai and Zou, 2013, 2015), which is the theoretical basis of the sure screening property. It assumes that 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 \( n > \max\{ Cq^2, 6^{1/(1-\kappa)} \} \)

\[
\Pr \left( \max_{1 \leq k \leq p} |\hat{\omega}_k^{\text{mRCC1}} - \omega_k^{\text{mRCC1}}| \geq c_2 q^4 n^{-\kappa} \right) \\
\leq p \cdot \left\{ 6q^2 \left( \exp(-c_3 n q^{-4}) + \exp(-c_4 n^3 q^{-4}) \right) \\
+ (2q^2 + 4q) \left( \exp(-c_5 n^{1-2\kappa}) + \exp(-c_6 n^{3-2\kappa}) \right) \right\}.
\]

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

\[
\Pr \left( \max_{1 \leq k \leq p} |\hat{\omega}_k^{\text{mRCC2}} - \omega_k^{\text{mRCC2}}| \geq c'_2 q^4 n^{-\kappa} \right) \\
\leq p \cdot \left\{ 6q^2 \exp(-c'_3 n q^{-4}) + (2q^2 + 4q) \exp(-c'_4 n^{1-2\kappa}) \right\}.
\]

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

\[
\Pr \left( \mathcal{A} \subset \hat{\mathcal{A}}_{t_n} \right) \geq 1 - p \cdot \left\{ 6q^2 \left( \exp(-c_3 n q^{-4}) + \exp(-c_4 n^3 q^{-4}) \right) \\
+ (2q^2 + 4q) \left( \exp(-c_5 n^{1-2\kappa}) + \exp(-c_6 n^{3-2\kappa}) \right) \right\}.
\]

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

\[
\Pr \left( \mathcal{A} \subset \hat{\mathcal{A}}'_{t'_n} \right) \geq 1 - p \cdot \left\{ 6q^2 \exp(-c'_3 n q^{-4}) + (2q^2 + 4q) \exp(-c'_4 n^{1-2\kappa}) \right\}.
\]
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 data dimensionality we can handle should satisfy \( \log(pq^2) = o(nq^{-4} + n^{1-2\kappa}) \) by both methods, with \( 0 < \kappa < 1/2 \). Under these settings we have the sure screening property \( \Pr(A \subset \hat{A}_{tn}) \to 1 \) or \( \Pr(A \subset \hat{A}'_{tn}) \to 1 \), respectively.

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

**Theorem 2.** Under Condition (C1), for any \( t_n = c_1q^4n^{-\kappa} \), there exist some positive constants \( c_3, c_4, c_5, c_6 \) and \( C \), such that when \( n > \max\{Cq^2, 6^{1/(1-\kappa)}\} \)

\[
\Pr\left(|\hat{A}_{tn}| \leq O(sq^{-2}n^\kappa)\right) \\
\geq 1 - p \cdot \left\{ 6q^2 \left( \exp(-c_3nq^{-4}) + \exp(-c_4n^3q^{-4}) \right) + (2q^2 + 4q) \left( \exp(-c_5n^{-2\kappa}) + \exp(-c_6n^{-3-2\kappa}) \right) \right\}.
\]

Under Condition (C1'), for any \( t'_n = c'_1q^4n^{-\kappa} \), there exist some positive
constants $c'_3$, $c'_4$, such that

$$
\Pr (|\hat{A}'_{n^*}| \leq O(sq^{-2}n^\kappa)) \\
\geq 1 - p \cdot \left\{ 6q^2 \exp(-c'_3 nq^{-4}) + (2q^2 + 4q) \exp(-c'_4 n^{1-2\kappa}) \right\}.
$$

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 0 exponentially fast.

5. Numerical Studies

In this section we evaluate the performance of all screening procedures discussed in this paper by 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 et al. (2019) and extend it to the multiple responses case in a similar way as mSIS does by aggregating the ranking statistics of the predictor with each response. We denote this method as mCCD. For the derivation and explanations of CCD, the readers are referred to Kong et al. (2019).

5.1 Simulations

We repeat each simulation 200 times and use the following three criteria adopted by Li et al. (2012b):

1. $S$: the minimum model size to include all the true predictors. We report the 5%, 25%, 50%, 75%, and 95% quantiles of $S$ out of 200 replications.
2. \( P_s \): the proportion that an individual true predictor is selected for a given model size \( d \) in the 200 replications.

3. \( P_a \): the proportion that all true predictors are selected for a given model size \( d \) in the 200 replications.

The \( S \) is used to measure the accuracy of a screening procedure. The smaller the \( S \) is, the less complex the resulting model is, and the better the screening procedure is. The \( P_s \) and \( P_a \) allow us to examine the chance that a screening procedure misses an individual predictor and all true predictors for a given model size \( d \), respectively. We present the simulation results of \( P_s, P_a \) with \( d = 2[n/\log n] \) for all the examples and the real data. We also tried \( d = [n/\log n] \) with quite similar outcomes and hence omit such results here for the sake of space.

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

\[
Y_j = 5X_1 + 5X_2 + 5X_3 + \epsilon_j, \quad j = 1, 2, \cdots, 10, \tag{5.1}
\]

The predictor vector \((X_1, \cdots, X_p)\) is drawn from a multivariate normal distribution \( \mathcal{N}(0, \Sigma) \), where \( \Sigma = CS(\rho) \) is a compound symmetric matrix with all the entries being \( \rho \) except for the diagonal elements being 1, and noise \( \epsilon_j \) follows the standard normal distributions. The sample size is \( n = 50 \), the number of the predictors and responses are \( p = 1000, q = 10 \) respectively,
and we consider \( \rho = 0, 0.1, 0.5, 0.9 \).

Table 1 summarizes the simulation results for \( S, P_s \) and \( P_a \). We can see that the mSIS works best in this example since this model is actually a univariate response data linear model with strong signal to noise ratio in every component. For the mCC and mRCC1, although the performance is acceptable but a little worse compared with other methods. One reason is that each response has the same strong signal which dominates the error term, hence the Pearson correlations among different pairs of the response are almost 1 (about 0.98, 0.99). In such case, the Condition (C1) for mRCC1 may be violated, and a similar reason goes for mCC since there is an inverse correlation matrix of the responses 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. (2012a):

\[
H(Y_j) = X_{10j-9} + X_{10j-8} + \epsilon_j, \quad j = 1, 2, \cdots, 10, \quad (5.2)
\]

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

\[
H(Y) = \frac{|Y|^{\lambda} \text{sgn}(Y) - 1}{\lambda}, \text{ when } \lambda = 0.25, 0.5, 0.75, 1; \quad H(Y) = \log Y, \text{ when } \lambda = 0.
\]
The variables \((X_1, \cdots, X_p)\) and noise \(\epsilon_j\) are generated in the same way that of Example 1. The number of the true variables is 20. \((n, p, q) = (200, 2000, 10)\) and \(\rho = 0.1, 0.5\), respectively.

The simulation results for \(S\) and \(P_a\) are reported in Table 2 and 3, respectively. We can see clearly that mRCC1 dramatically outperforms other methods especially when \(\rho = 0.5\), and the results are almost invariant under transformations (a little difference due to different random errors generated for models with different \(\lambda\)). 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 (\(\lambda\) decrease from 1), the performance of mSIS, mNIS, DCS and mCC quickly deteriorates due to the existence of the nonlinearity and heavy-tailed responses.

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

\[
Y_j = 2 \sin (\alpha_{j1}X_1 + \alpha_{j2}X_2 + \alpha_{j3}X_3 + \alpha_{j4}X_4 + \alpha_{j5}X_5) + \epsilon_j, \quad j = 1, 2, \cdots, 20, \tag{5.3}
\]

where \(\alpha_{j1}, \cdots, \alpha_{j5} \sim \text{Unif}(0,1)\) independently for \(j = 1, \cdots, 20\). Once the parameter is drawn, the model is fixed. We generate \((X_1, \cdots, X_p)\) and noise \(\epsilon_j\) same as Example 1. \((n, p, q) = (200, 2000, 20)\) and \(\rho = 0.5, 0.8\), respectively.
Table 4 gives the results for $S$. Table 5 shows the results for $P_s$ and $P_a$. For this nonlinear model, mRCC1 and mRCC2 are still robust and encouraging and they perform best.

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

$$Y_j = 4X_{jk_1} + 2\tan(\pi X_{jk_2}/2) + 5X_{jk_3}^2 + \epsilon_j, \quad j = 1, 2, \cdots, 20.$$  

(5.4)

The predictors follow Unif(0,1) independently and $\epsilon_j$’s follow $N(0, 1)$ which are independent of predictors. For each $j$, the indexes $\{k_1, k_2, k_3\}$ are randomly drawn from $\{1, 2, \cdots, 10\}$. Once the indexes are drawn, the model is fixed. In our simulation, we checked that $X_1, X_2, \cdots, X_{10}$ are all included in the model, hence the number of true predictors is 10. The sample size is $n = 200$, the number of the predictors and responses are $p = 2000, q = 20$ respectively.

From Table 6 and Table 7, we see that the mRCC1 and mRCC2 achieve perfect selection with the oracle variables, although this is a nonlinear model and heavy-tailed data exists. The performance of DCS apparently falls behind.

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

$$Y_j \sim \text{Poisson}(\mu_j), \quad \mu_j = \exp(0.8X_1 - 0.8X_2), \quad j = 1, \cdots, 10,$$
where predictors $X_k \sim t_4$ independently, for $k = 1, 2, \cdots, 2000$. The sample size is 200 and $q = 10$.

The results are 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 $\tau$ correlation in mRRCS and mRCC2 uses formula (2.4) in Li et al. (2012a) which may be inadequate for tied variables).

5.2 Genomic Data Example

The breast cancer dataset was detailed described by Chin et al. (2006) and analyzed by Witten et al. (2009), Chen et al. (2013) and Molstad and Rothman (2016). The dataset is publicly available in the R package PMA (Witten et al., 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, since certain types of cancer are characterized by unusual DNA copy-number changes which has been revealed by 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 the aforementioned screening methods to carry out the multivariate response regression for 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 top \( d = 2[n_{\text{train}} / \log n_{\text{train}}] \) genes where \( n_{\text{train}} \) 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 prediction in the test samples. The model fitting process and tuning parameter selection are implemented using R package \textit{glmnet}. Following Chen et al. (2013), we calculate the mean squared prediction error \( \| Y_{\text{test}} - X_{\text{test}} \hat{B} \|_F^2 / \gamma_{\text{test}} \), where \((Y_{\text{test}}, X_{\text{test}})\) denotes the test set, \( \hat{B} \) denotes the estimated coefficient matrix and \( n_{\text{test}} \) is the sample size of test set. The above procedure is repeated 200 times.

The means of the prediction errors with their standard errors are presented in Table 9. For each splitting ratio, the mRCC2 enjoys the outstanding predictive performance. To check whether or not the MSE’s for the proposed approaches are significantly different than those for the other methods, we perform two-sided paired samples t-tests for the mCC, mRCC1, mRCC2 against other methods and the corresponding p-values are
presented in Table 10. We also conduct a one-sided paired samples $t$-test for mRCC2 only and its $p$-values are still 0, which confirms that mRCC2 has significantly lower prediction errors than others. Furthermore, we list genes with top 7 highest selection frequency by mRCC2 in Table 11. We can see top 3 among them 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 profile, which could be very useful in subsequent studies.
Acknowledgements We sincerely thank the editor, AE and referees for their helpful comments. He is supported by the Youth Program of National Natural Science Foundation of China (11901289). Zou’s work is supported in part by NSF DMS 1915842.

A. Appendix

Lemma 1 (Theorem A, Serfling (1980), p201). Let $X_1, X_2, \cdots, X_n$ be independent observations on a distribution function $F$. Let $h = h(x_1, \cdots, x_m)$ be a kernel for a “parametric function” $\theta = \theta(F)$, with $a \leq h(x_1, \cdots, x_m) \leq b$.

Put $\theta = E\{h(X_1, \cdots, X_m)\}$, then, for $t > 0$ and $n \geq m$,

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

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

$$U_n = \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, \cdots, X_{i_m})$$

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

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

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

for $n > 3$, where $r_s$ is the population Spearman correlation.
Table 1: The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $S$ and the proportions of $P_s$ and $P_a$ out of 200 replications in Example 1

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>Method</th>
<th>$S$</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
<th>$P_s$</th>
<th>$P_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>mSIS</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>5.0</td>
<td>16.1</td>
<td></td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>mNIS</td>
<td>3.0</td>
<td>4.0</td>
<td>8.0</td>
<td>30.3</td>
<td>144.8</td>
<td></td>
<td>0.92</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>mRRCS</td>
<td>3.0</td>
<td>3.0</td>
<td>4.0</td>
<td>7.0</td>
<td>33.2</td>
<td></td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>DCS</td>
<td>3.0</td>
<td>3.0</td>
<td>4.0</td>
<td>7.0</td>
<td>27.1</td>
<td></td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>mCCD</td>
<td>3.0</td>
<td>3.0</td>
<td>4.0</td>
<td>7.0</td>
<td>24.2</td>
<td></td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>mCC</td>
<td>4.0</td>
<td>11.0</td>
<td>25.0</td>
<td>100.8</td>
<td>283.2</td>
<td>0.83</td>
<td>0.85</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>mRCC1</td>
<td>4.0</td>
<td>15.8</td>
<td>39.5</td>
<td>130.5</td>
<td>427.0</td>
<td>0.75</td>
<td>0.81</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>mRCC2</td>
<td>3.0</td>
<td>3.0</td>
<td>4.0</td>
<td>9.0</td>
<td>40.1</td>
<td></td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>0.1</td>
<td>mSIS</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>5.0</td>
<td>23.1</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>mNIS</td>
<td>3.0</td>
<td>3.0</td>
<td>6.0</td>
<td>12.0</td>
<td>41.4</td>
<td></td>
<td>0.86</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>mRRCS</td>
<td>3.0</td>
<td>3.0</td>
<td>4.0</td>
<td>8.0</td>
<td>41.2</td>
<td></td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>DCS</td>
<td>3.0</td>
<td>3.0</td>
<td>4.0</td>
<td>7.0</td>
<td>35.1</td>
<td></td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>mCCD</td>
<td>3.0</td>
<td>3.0</td>
<td>4.0</td>
<td>7.0</td>
<td>34.0</td>
<td></td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>mCC</td>
<td>3.0</td>
<td>6.0</td>
<td>16.0</td>
<td>51.3</td>
<td>144.6</td>
<td>0.76</td>
<td>0.83</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>mRCC1</td>
<td>4.0</td>
<td>9.0</td>
<td>24.5</td>
<td>67.0</td>
<td>226.9</td>
<td>0.74</td>
<td>0.80</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>mRCC2</td>
<td>3.0</td>
<td>3.0</td>
<td>4.0</td>
<td>9.0</td>
<td>44.1</td>
<td></td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>0.5</td>
<td>mSIS</td>
<td>3.0</td>
<td>4.0</td>
<td>6.5</td>
<td>18.3</td>
<td>100.4</td>
<td></td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>mNIS</td>
<td>3.0</td>
<td>5.0</td>
<td>9.0</td>
<td>27.0</td>
<td>124.2</td>
<td></td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>mRRCS</td>
<td>3.0</td>
<td>6.0</td>
<td>13.0</td>
<td>35.3</td>
<td>173.4</td>
<td>0.98</td>
<td>0.97</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>DCS</td>
<td>3.0</td>
<td>5.0</td>
<td>11.0</td>
<td>29.0</td>
<td>141.2</td>
<td>0.98</td>
<td>0.97</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>mCCD</td>
<td>3.0</td>
<td>7.0</td>
<td>14.5</td>
<td>40.5</td>
<td>192.7</td>
<td>0.99</td>
<td>0.97</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>mCC</td>
<td>3.0</td>
<td>6.0</td>
<td>14.0</td>
<td>41.0</td>
<td>208.5</td>
<td>0.78</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>mRCC1</td>
<td>4.0</td>
<td>11.8</td>
<td>38.0</td>
<td>83.0</td>
<td>371.1</td>
<td>0.68</td>
<td>0.83</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>mRCC2</td>
<td>3.0</td>
<td>6.0</td>
<td>14.0</td>
<td>35.3</td>
<td>186.6</td>
<td>0.98</td>
<td>0.97</td>
<td>0.87</td>
</tr>
<tr>
<td>0.9</td>
<td>mSIS</td>
<td>3.0</td>
<td>4.0</td>
<td>8.0</td>
<td>19.8</td>
<td>91.1</td>
<td></td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>mNIS</td>
<td>3.0</td>
<td>5.0</td>
<td>11.0</td>
<td>28.0</td>
<td>105.1</td>
<td></td>
<td>0.71</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>mRRCS</td>
<td>4.0</td>
<td>13.0</td>
<td>36.0</td>
<td>77.5</td>
<td>231.0</td>
<td>0.91</td>
<td>0.91</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>DCS</td>
<td>3.0</td>
<td>6.0</td>
<td>16.0</td>
<td>39.3</td>
<td>142.5</td>
<td>0.94</td>
<td>0.93</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>mCCD</td>
<td>5.0</td>
<td>34.8</td>
<td>75.0</td>
<td>173.5</td>
<td>424.4</td>
<td>0.95</td>
<td>0.93</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>mCC</td>
<td>3.0</td>
<td>7.0</td>
<td>14.5</td>
<td>38.3</td>
<td>166.3</td>
<td>0.50</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>mRCC1</td>
<td>8.0</td>
<td>25.0</td>
<td>65.0</td>
<td>144.5</td>
<td>437.5</td>
<td>0.36</td>
<td>0.50</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>mRCC2</td>
<td>4.0</td>
<td>13.0</td>
<td>36.0</td>
<td>77.0</td>
<td>258.7</td>
<td>0.91</td>
<td>0.92</td>
<td>0.67</td>
</tr>
</tbody>
</table>
Table 2: The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $S$ out of 200 replications in Example 2

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Method</th>
<th>CS(0.1)</th>
<th>CS(0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>mSIS</td>
<td>202.7 421.0 653.0 940.3 1413.7</td>
<td>1237.9 1553.8 1723.5 1829.0 1945.3</td>
</tr>
<tr>
<td></td>
<td>mNIS</td>
<td>600.6 880.5 1118.5 1310.5 1655.4</td>
<td>1296.9 1563.3 1729.0 1876.3 1955.2</td>
</tr>
<tr>
<td></td>
<td>mRRCS</td>
<td>20.0 23.0 28.5 43.0 76.3</td>
<td>74.5 1054.5 1336.5 1547.5 1816.5</td>
</tr>
<tr>
<td></td>
<td>DCS</td>
<td>423.4 712.8 924.5 1194.0 1665.2</td>
<td>1143.0 1485.3 1685.0 1825.0 1957.3</td>
</tr>
<tr>
<td></td>
<td>mCC</td>
<td>137.0 294.3 549.0 869.3 1440.7</td>
<td>905.0 1251.5 1440.5 1655.3 1861.4</td>
</tr>
<tr>
<td></td>
<td>mRCC1</td>
<td>20.0 20.0 20.0 20.0 21.1</td>
<td>40.0 96.0 155.5 313.0 731.2</td>
</tr>
<tr>
<td></td>
<td>mRCC2</td>
<td>20.0 20.0 20.0 20.0 24.0</td>
<td>95.9 179.0 288.5 513.0 1012.1</td>
</tr>
<tr>
<td>0.25</td>
<td>mSIS</td>
<td>30.0 46.0 71.5 138.3 397.1</td>
<td>1128.0 1363.0 1598.0 1767.3 1923.7</td>
</tr>
<tr>
<td></td>
<td>mNIS</td>
<td>141.0 254.0 346.0 485.0 827.5</td>
<td>1078.9 1432.3 1606.0 1774.3 1919.3</td>
</tr>
<tr>
<td></td>
<td>mRRCS</td>
<td>20.0 23.0 28.0 43.3 101.2</td>
<td>733.9 1027.0 1277.0 1542.8 1786.0</td>
</tr>
<tr>
<td></td>
<td>DCS</td>
<td>56.0 103.8 168.0 268.8 601.3</td>
<td>981.8 1305.5 1696.3 1925.5 2195.2</td>
</tr>
<tr>
<td></td>
<td>mCC</td>
<td>20.0 21.0 24.0 34.3 84.1</td>
<td>360.9 591.8 827.0 1108.3 1621.0</td>
</tr>
<tr>
<td></td>
<td>mRCC1</td>
<td>20.0 20.0 20.0 20.0 22.0</td>
<td>45.8 90.5 151.0 292.0 693.3</td>
</tr>
<tr>
<td></td>
<td>mRCC2</td>
<td>20.0 20.0 20.0 20.0 24.0</td>
<td>92.0 192.5 304.5 685.0 1634.1</td>
</tr>
<tr>
<td>0.5</td>
<td>mSIS</td>
<td>22.0 27.8 41.0 65.8 173.0</td>
<td>942.0 1214.5 1453.5 1654.0 1888.2</td>
</tr>
<tr>
<td></td>
<td>mNIS</td>
<td>59.0 96.5 151.0 237.0 503.3</td>
<td>1056.0 1310.0 1498.0 1686.8 1885.1</td>
</tr>
<tr>
<td></td>
<td>mRRCS</td>
<td>20.0 22.0 29.0 43.0 105.1</td>
<td>720.6 1044.8 1322.5 1509.5 1826.1</td>
</tr>
<tr>
<td></td>
<td>DCS</td>
<td>25.0 36.8 54.0 99.0 235.2</td>
<td>884.2 1191.3 1409.0 1606.3 1886.6</td>
</tr>
<tr>
<td></td>
<td>mCC</td>
<td>22.0 25.0 36.0 58.0 129.4</td>
<td>360.9 591.8 827.0 1108.3 1621.0</td>
</tr>
<tr>
<td></td>
<td>mRCC1</td>
<td>20.0 20.0 20.0 20.0 22.0</td>
<td>45.8 90.5 151.0 292.0 693.3</td>
</tr>
<tr>
<td></td>
<td>mRCC2</td>
<td>20.0 20.0 20.0 20.0 24.0</td>
<td>92.0 192.5 304.5 685.0 1634.1</td>
</tr>
<tr>
<td>0.75</td>
<td>mSIS</td>
<td>21.0 23.0 33.0 51.0 106.6</td>
<td>795.0 1122.0 1390.0 1625.5 1879.0</td>
</tr>
<tr>
<td></td>
<td>mNIS</td>
<td>27.0 43.0 66.5 118.3 301.5</td>
<td>845.6 1132.0 1420.0 1613.5 1859.7</td>
</tr>
<tr>
<td></td>
<td>mRRCS</td>
<td>21.0 23.8 29.0 43.0 100.5</td>
<td>740.8 996.5 1273.0 1547.3 1822.3</td>
</tr>
<tr>
<td></td>
<td>DCS</td>
<td>21.0 25.0 32.5 50.3 117.1</td>
<td>766.7 1079.5 1325.0 1593.8 1821.3</td>
</tr>
<tr>
<td></td>
<td>mCC</td>
<td>21.0 27.0 36.0 57.0 130.1</td>
<td>906.5 1172.3 1415.0 1622.5 1889.5</td>
</tr>
<tr>
<td></td>
<td>mRCC1</td>
<td>20.0 20.0 20.0 20.0 22.0</td>
<td>45.0 87.8 175.0 310.5 832.1</td>
</tr>
<tr>
<td></td>
<td>mRCC2</td>
<td>20.0 20.0 20.0 20.0 23.1</td>
<td>84.8 197.3 330.0 502.5 1097.3</td>
</tr>
<tr>
<td>1</td>
<td>mSIS</td>
<td>21.0 24.0 29.5 43.0 89.7</td>
<td>661.9 1126.0 1358.0 1616.8 1841.0</td>
</tr>
<tr>
<td></td>
<td>mNIS</td>
<td>24.0 38.0 59.0 109.3 228.2</td>
<td>686.7 1170.0 1391.0 1596.5 1862.1</td>
</tr>
<tr>
<td></td>
<td>mRRCS</td>
<td>20.0 22.8 28.0 42.0 99.1</td>
<td>649.1 1064.5 1322.0 1542.8 1812.1</td>
</tr>
<tr>
<td></td>
<td>DCS</td>
<td>21.0 24.0 30.5 44.0 95.1</td>
<td>702.6 1099.5 1351.5 1576.0 1809.6</td>
</tr>
<tr>
<td></td>
<td>mCC</td>
<td>22.0 26.0 34.0 56.5 120.4</td>
<td>776.7 1174.0 1427.5 1651.8 1845.1</td>
</tr>
<tr>
<td></td>
<td>mRCC1</td>
<td>20.0 20.0 20.0 20.0 22.1</td>
<td>47.9 86.5 157.5 308.5 754.3</td>
</tr>
<tr>
<td></td>
<td>mRCC2</td>
<td>20.0 20.0 20.0 20.0 25.0</td>
<td>92.4 189.0 319.0 533.3 1144.2</td>
</tr>
</tbody>
</table>
Table 3: The proportions of $\mathcal{P}_a$ in Example 2

<table>
<thead>
<tr>
<th>Method</th>
<th>$\lambda = 0$</th>
<th>$\lambda = 0.25$</th>
<th>$\lambda = 0.5$</th>
<th>$\lambda = 0.75$</th>
<th>$\lambda = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mSIS</td>
<td>0.00</td>
<td>0.52</td>
<td>0.80</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>mNIS</td>
<td>0.00</td>
<td>0.00</td>
<td>0.14</td>
<td>0.57</td>
<td>0.59</td>
</tr>
<tr>
<td>mRRCS</td>
<td>0.95</td>
<td>0.94</td>
<td>0.90</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td>DCS</td>
<td>0.00</td>
<td>0.14</td>
<td>0.63</td>
<td>0.87</td>
<td>0.91</td>
</tr>
<tr>
<td>mCCD</td>
<td>0.58</td>
<td>0.83</td>
<td>0.84</td>
<td>0.84</td>
<td>0.85</td>
</tr>
<tr>
<td>mCC</td>
<td>0.02</td>
<td>0.94</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>mRCC1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>mRCC2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>CS(0.5)</th>
<th>CS(0.8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>25%</td>
</tr>
<tr>
<td>mSIS</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>mNIS</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>mRRCS</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>DCS</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>mCCD</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>mCC</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>mRCC1</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>mRCC2</td>
<td>5.0</td>
<td>5.0</td>
</tr>
</tbody>
</table>
Table 5: The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 3

<table>
<thead>
<tr>
<th>Method</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>All</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>mSIS</td>
<td>0.86</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.86</td>
<td>0.27</td>
<td>0.95</td>
<td>0.83</td>
<td>0.94</td>
<td>0.82</td>
<td>0.17</td>
</tr>
<tr>
<td>mNIS</td>
<td>0.93</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.93</td>
<td>0.69</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.66</td>
</tr>
<tr>
<td>mRRCS</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.96</td>
<td>0.57</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>0.54</td>
</tr>
<tr>
<td>DCS</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.96</td>
<td>0.59</td>
<td>0.98</td>
<td>0.97</td>
<td>1.00</td>
<td>1.00</td>
<td>0.54</td>
</tr>
<tr>
<td>mCCD</td>
<td>0.92</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.92</td>
<td>0.46</td>
<td>0.97</td>
<td>0.96</td>
<td>1.00</td>
<td>0.99</td>
<td>0.42</td>
</tr>
<tr>
<td>mCC</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>mRCC1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>mRCC2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>mSIS</td>
<td>908.2</td>
<td>1335.0</td>
<td>1573.0</td>
<td>1794.3</td>
<td>1960.1</td>
</tr>
<tr>
<td>mNIS</td>
<td>1085.4</td>
<td>1479.8</td>
<td>1665.5</td>
<td>1835.5</td>
<td>1965.2</td>
</tr>
<tr>
<td>mRRCS</td>
<td>10.0</td>
<td>11.0</td>
<td>13.0</td>
<td>25.0</td>
<td>76.0</td>
</tr>
<tr>
<td>DCS</td>
<td>275.6</td>
<td>600.0</td>
<td>942.0</td>
<td>1375.5</td>
<td>1806.4</td>
</tr>
<tr>
<td>mCCD</td>
<td>15.0</td>
<td>56.8</td>
<td>179.5</td>
<td>368.5</td>
<td>807.4</td>
</tr>
<tr>
<td>mCC</td>
<td>10.0</td>
<td>10.0</td>
<td>17.0</td>
<td>141.3</td>
<td>912.1</td>
</tr>
<tr>
<td>mRCC1</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>mRCC2</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
</tr>
</tbody>
</table>
Table 7: The proportions of $P_s$ and $P_a$ in Example 4

<table>
<thead>
<tr>
<th>Method</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>mSIS</td>
<td>0.10</td>
<td>0.03</td>
<td>0.90</td>
<td>0.09</td>
<td>0.07</td>
<td>0.92</td>
<td>0.07</td>
<td>0.78</td>
<td>0.97</td>
<td>0.88</td>
<td>0.00</td>
</tr>
<tr>
<td>mNIS</td>
<td>0.04</td>
<td>0.02</td>
<td>0.99</td>
<td>0.05</td>
<td>0.04</td>
<td>1.00</td>
<td>0.02</td>
<td>0.92</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>mRRCS</td>
<td>1.00</td>
<td>0.95</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>DCS</td>
<td>0.49</td>
<td>0.16</td>
<td>0.91</td>
<td>0.42</td>
<td>0.37</td>
<td>0.94</td>
<td>0.24</td>
<td>0.93</td>
<td>0.96</td>
<td>0.90</td>
<td>0.00</td>
</tr>
<tr>
<td>mCCD</td>
<td>1.00</td>
<td>0.38</td>
<td>1.00</td>
<td>0.95</td>
<td>0.98</td>
<td>1.00</td>
<td>0.78</td>
<td>1.00</td>
<td>1.00</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>mCC</td>
<td>1.00</td>
<td>1.00</td>
<td>0.82</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.81</td>
<td>0.65</td>
</tr>
<tr>
<td>mRCC1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>mRCC2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 8: The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $S$ and the proportions of $P_s$ and $P_a$ out of 200 replications in Example 5

<table>
<thead>
<tr>
<th>Method</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
<th>X1</th>
<th>X2</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>mSIS</td>
<td>2.0</td>
<td>3.0</td>
<td>15.5</td>
<td>136.0</td>
<td>1346.5</td>
<td>0.85</td>
<td>0.86</td>
<td>0.70</td>
</tr>
<tr>
<td>mNIS</td>
<td>3.0</td>
<td>27.0</td>
<td>58.5</td>
<td>223.0</td>
<td>1405.2</td>
<td>0.79</td>
<td>0.80</td>
<td>0.59</td>
</tr>
<tr>
<td>mRRCS</td>
<td>1079.8</td>
<td>1848.3</td>
<td>1971.0</td>
<td>1999.0</td>
<td>2000.0</td>
<td>0.01</td>
<td>1.00</td>
<td>0.01</td>
</tr>
<tr>
<td>DCS</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>5.3</td>
<td>481.1</td>
<td>0.92</td>
<td>0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>mCCD</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>mCC</td>
<td>2.0</td>
<td>6.0</td>
<td>16.0</td>
<td>47.5</td>
<td>664.9</td>
<td>0.88</td>
<td>0.92</td>
<td>0.80</td>
</tr>
<tr>
<td>mRCC1</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>mRCC2</td>
<td>1208.4</td>
<td>1882.5</td>
<td>1980.0</td>
<td>1999.3</td>
<td>2000.0</td>
<td>0.01</td>
<td>1.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Table 9: Means of the prediction errors in 200 times randomly split genomic data. The standard errors of prediction errors are shown in the parenthesis, where $\gamma$ is the proportion of the training samples. NM is the null model using componentwise means of training responses to predict test sample.

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>mSIS</th>
<th>mNIS</th>
<th>mRRCS</th>
<th>DCS</th>
<th>mCCD</th>
<th>mCC</th>
<th>mRCC1</th>
<th>mRCC2</th>
<th>NM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.769</td>
<td>0.758</td>
<td>0.791</td>
<td>0.765</td>
<td>0.781</td>
<td>0.786</td>
<td>0.774</td>
<td>0.738</td>
<td>1.021</td>
</tr>
<tr>
<td></td>
<td>(0.066)</td>
<td>(0.068)</td>
<td>(0.065)</td>
<td>(0.068)</td>
<td>(0.067)</td>
<td>(0.08)</td>
<td>(0.079)</td>
<td>(0.067)</td>
<td>(0.089)</td>
</tr>
<tr>
<td>0.8</td>
<td>0.737</td>
<td>0.737</td>
<td>0.784</td>
<td>0.735</td>
<td>0.753</td>
<td>0.756</td>
<td>0.756</td>
<td>0.685</td>
<td>1.047</td>
</tr>
<tr>
<td></td>
<td>(0.135)</td>
<td>(0.129)</td>
<td>(0.141)</td>
<td>(0.129)</td>
<td>(0.136)</td>
<td>(0.132)</td>
<td>(0.146)</td>
<td>(0.117)</td>
<td>(0.184)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>SIS</th>
<th>NIS</th>
<th>Kendall</th>
<th>DC</th>
<th>mCCD</th>
<th>mCC</th>
<th>mRCC1</th>
<th>mRCC2</th>
<th>NM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>mCC</td>
<td>0.001</td>
<td>0.307</td>
<td>0</td>
<td>0.336</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mRCC1</td>
<td>0.286</td>
<td>0</td>
<td>0.001</td>
<td>0.072</td>
<td>0.201</td>
<td>0.01</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mRCC2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>0.8</td>
<td>mCC</td>
<td>0.081</td>
<td>0.079</td>
<td>0.016</td>
<td>0.058</td>
<td>0.743</td>
<td>-</td>
<td>0.917</td>
<td>0</td>
</tr>
<tr>
<td>mRCC1</td>
<td>0.094</td>
<td>0.092</td>
<td>0.015</td>
<td>0.072</td>
<td>0.785</td>
<td>0.917</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mRCC2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 11: Genes with top 7 highest selection frequency by mRCC2.

<table>
<thead>
<tr>
<th>$\gamma$ = 0.5</th>
<th>genenames</th>
<th>COX4I1</th>
<th>KIAA0174</th>
<th>FLJ13868</th>
<th>KIAA1007</th>
<th>USP10</th>
<th>PARN</th>
<th>KATNB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>0.82</td>
<td>0.76</td>
<td>0.72</td>
<td>0.58</td>
<td>0.57</td>
<td>0.56</td>
<td>0.55</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\gamma$ = 0.8</th>
<th>genenames</th>
<th>COX4I1</th>
<th>FLJ13868</th>
<th>KIAA0174</th>
<th>SF3B3</th>
<th>KATNB1</th>
<th>KIAA1007</th>
<th>USP10</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.97</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>
Proof of Lemma 2. If we take \( h \) in (6.1) to be the kernel of degree \( m = 2 \) given by
\[
h_{\tilde{\tau}} \left( \left( x_1, y_1 \right), \left( x_2, y_2 \right) \right) = \text{sgn}(x_1 - x_2) \text{sgn}(y_1 - y_2),
\]
then \( \tilde{\tau} \triangleq U_{h_{\tilde{\tau}}} \) is the sample Kendall’s tau correlation. If we take \( h \) in (6.1) to be the kernel of degree \( m = 3 \) given by
\[
h_{\tilde{r}_s} \left( \left( x_1, y_1 \right), \left( x_2, y_2 \right), \left( x_3, y_3 \right) \right) = \frac{1}{2} \sum_{i,j,l=1}^{3} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_l),
\]
and define \( \tilde{r}_s \triangleq U_{h_{\tilde{r}_s}} \) \cite{Hoeffding}, showed that
\[
\hat{\tau} = \frac{n - 2}{n + 1} \tilde{r}_s + \frac{3}{n + 1} \hat{\tau}. \tag{6.2}
\]
Hence, the dominating term \( \tilde{r}_s \) of Spearman correlation is a U-statistic.

Since \( \tilde{r}_s, \hat{\tau} \) are unbiased estimators of their population version \( r_s, \tau \) respectively, and \(|h_{\tilde{r}_s}| \leq 1, |h_{\hat{\tau}}| \leq 1\), by Lemma 1.

\[
\Pr \left( \tilde{r}_s - r_s \geq \frac{\delta}{2} \right) \leq \exp \left( -\frac{(n - 3)\delta^2}{24} \right), \tag{6.3}
\]
\[
\Pr \left( \hat{\tau} - \tau \geq \frac{\delta}{2} \right) \leq \exp \left( -\frac{(n - 2)\delta^2}{16} \right), \tag{6.4}
\]
On Sure Screening with Multiple Responses

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

$$\Pr \left( |\hat{r}_s - r_s| \geq \frac{6}{n + \delta} \right) \leq \Pr \left( \hat{r}_s - r_s \geq \frac{6}{n + 1} + \delta \right) + \Pr \left( \hat{r}_s - r_s \leq -\frac{6}{n + 1} - \delta \right)$$

$$\leq \Pr \left( \hat{r}_s - r_s - \frac{3(\tau - r_s)}{n + 1} \geq \delta \right) + \Pr \left( \hat{r}_s - r_s - \frac{3(\tau - r_s)}{n + 1} \leq -\delta \right)$$

$$\leq \Pr \left( \frac{n - 2}{n + 1}(\hat{r}_s - r_s) + \frac{3}{n + 1}(\hat{r}_s - r_s) \geq \delta \right) + \Pr \left( \frac{n - 2}{n + 1}(\hat{r}_s - r_s) + \frac{3}{n + 1}(\hat{r}_s - r_s) \leq -\delta \right)$$

$$\leq 2 \exp \left( -\frac{(n - 3)(n + 1)^2 \delta^2}{24(n - 2)^2} \right) + 2 \exp \left( -\frac{(n - 2)(n + 1)^2 \delta^2}{16} \right).$$

Throughout the rest of the paper, for any matrix $A$, denote $\|A\| = \sqrt{\lambda_{\max}(A^T A)}$ be the spectral norm and $\|A\|_{\max} = \max_{i,j} |A_{i,j}|$ 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 $n > Cq^2$

$$\Pr \left( \|\Sigma_{R(Y)}^{-1} - \Sigma_{R(Y)}^{-1}\| \geq c_8 \|\Sigma_{R(Y)}^{-1}\| \right) \leq 2q^2 \exp(-c_3 nq^{-4}) + 2q^2 \exp(-c_4 n^3 q^{-4}).$$

**Proof of Lemma 3.** For any symmetric matrices $A$, $B$ and $D$, by the similar argument in the proof of Lemma 5 of [Fan et al. (2011)](#), we have

$$|\lambda_{\min}(A) - \lambda_{\min}(B)| \leq \max\{|\lambda_{\min}(A - B)|, |\lambda_{\min}(B - A)|\},$$
\[ |\lambda_{\min}(D)| \leq d\|D\|_{\max}, \quad |\lambda_{\min}(-D)| \leq d\|D\|_{\max}, \]

where \( d \) is the dimension of \( D \). Hence,

\[ |\lambda_{\min}((\hat{\Sigma}_R(Y)) - \lambda_{\min}(\Sigma_R(Y))| \leq q\|\hat{\Sigma}_R(Y) - \Sigma_R(Y)\|_{\max}. \]

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

\[
\Pr \left( |\lambda_{\min}((\hat{\Sigma}_R(Y)) - \lambda_{\min}(\Sigma_R(Y))| \geq q\left( \frac{6}{n} + \delta_1 \right) \right)
\leq q^2 \Pr \left( |\hat{r}_s(Y_j, Y_l) - r_s(Y_j, Y_l)| \geq \frac{6}{n} + \delta_1 \right)
\leq 2q^2 \exp(-\tilde{c}_3 n \delta_1^2) + 2q^2 \exp(-\tilde{c}_4 n^3 \delta_1^2),
\]

for some positive constant \( \tilde{c}_3 \) and \( \tilde{c}_4 \). Take \( \delta_1 = c_9 c_0 q^{-2} - \frac{6}{n} \) in the above, where \( c_9 \in (0, 1) \), denote \( C = 6/(c_9 c_0) \), by the Condition (C1), when \( n > Cq^2 \) it follows that

\[
\Pr \left( |\lambda_{\min}((\hat{\Sigma}_R(Y)) - \lambda_{\min}(\Sigma_R(Y))| \geq c_9 \lambda_{\min}(\Sigma_R(Y)) \right)
\leq 2q^2 \exp(-c_3 n \delta_1^2) + 2q^2 \exp(-c_4 n^3 \delta_1^2),
\]

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 et al. (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 \( \lambda_{\min}^{-1}(D) = \lambda_{\max}(D^{-1}) \),
we have

\[
\text{Pr} \left( \left| \| \hat{\Sigma}_R^{-1}(Y) \| - \| \Sigma_R^{-1}(Y) \| \right| \geq c_8 \| \Sigma_R^{-1}(Y) \| \right)
\]

\[
= \text{Pr} \left( |\lambda_{\min}(\hat{\Sigma}_R(Y)) - \lambda_{\min}(\Sigma_R(Y))| \geq c_8 \lambda_{\min}(\Sigma_R(Y)) \right)
\]

\[
\leq \text{Pr} \left( |\lambda_{\min}(\hat{\Sigma}_R(Y)) - \lambda_{\min}(\Sigma_R(Y))| \geq c_9 \lambda_{\min}(\Sigma_R(Y)) \right)
\]

\[
\leq 2q^2 \exp(-c_3 n q^{-d}) + 2q^2 \exp(-c_4 n^3 q^{-d}),
\]

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

\[\square\]

**Proof of Theorem** We only focus on the proof for mRCC1, since the proof for mRCC2 is similar by modifying tail probability using \( (6.4) \) in Lemma and the following. For the first part of the theorem, recall that

\[
\hat{\omega}_k^{\text{mRCC1}} = \hat{r}_k^T \hat{\Sigma}_R^{-1}(Y) \hat{r}_k,
\]

\[
\omega_k^{\text{mRCC1}} = r_k^T \Sigma_R^{-1}(Y) r_k,
\]

we have

\[
\hat{\omega}_k^{\text{mRCC1}} - \omega_k^{\text{mRCC1}} = (\hat{r}_k - r_k)^T \hat{\Sigma}_R^{-1}(Y) (\hat{r}_k - r_k)
\]

\[+ 2(\hat{r}_k - r_k)^T \hat{\Sigma}_R^{-1}(Y) r_k
\]

\[+ r_k^T (\hat{\Sigma}_R(Y) - \Sigma_R^{-1}(Y)) r_k
\]

\[\triangleq I_1 + I_2 + I_3.
\]
Note that
\[ I_1 \leq \| \hat{\Sigma}_{R(Y)}^{-1} \| \cdot \| \hat{r}_k - r_k \|^2. \]

By Lemma 2, for any \( \delta > 0 \), the union bound of probability is
\[
\Pr \left( \| \hat{r}_k - r_k \|^2 \geq q \left( \frac{6}{n} + \delta \right)^2 \right) \leq q \Pr \left( \| \hat{r}_s(X_k, Y_j) - r_s(X_k, Y_j) \|^2 > \left( \frac{6}{n} + \delta \right)^2 \right)
\leq 2q \exp(-\tilde{c}_5 n \delta^2) + 2q \exp(-\tilde{c}_6 n^3 \delta^2),
\]
for some positive constant \( \tilde{c}_5 \) and \( \tilde{c}_6 \). Under Condition (C1),
\[
\| \Sigma_{R(Y)}^{-1} \| \leq c_0^{-1} q.
\]

By Lemma 3,
\[
\Pr \left( \| \hat{\Sigma}_{R(Y)}^{-1} \| \geq (c_8 + 1)c_0^{-1} q \right) \leq 2q^2 \exp(-c_3 n q^{-4}) + 2q^2 \exp(-c_4 n^3 q^{-4}),
\]
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
\[
\Pr \left( |I_1| \geq (c_8 + 1)c_0^{-1} q^2 \left( \frac{6}{n} + \delta \right)^2 \right) \leq 2q^2 \exp(-c_3 n q^{-4}) + 2q^2 \exp(-c_4 n^3 q^{-4})
+ 2q \exp(-\tilde{c}_5 n \delta^2) + 2q \exp(-\tilde{c}_6 n^3 \delta^2).
\]

We next deal with the probability bound for \( I_2 \). Note that
\[
|I_2| \leq 2\| (\hat{r}_k - r_k)^T \| \cdot \| \hat{\Sigma}_{R(Y)}^{-1} \| \cdot \| r_k \|. 
\]

It is obvious that
\[
\| r_k \|^2 \leq q.
\]
Hence, the union bound of probability for $I_2$ is

$$\Pr \left( |I_2| \geq 2(c_8 + 1)c_0^{-1}q^2\left(\frac{6}{n} + \delta\right) \right) \leq 2q^2 \exp(-c_3nq^{-4}) + 2q^2 \exp(-c_4n^3q^{-4})$$

$$+ 2q \exp(-\tilde{c}_5n\delta^2) + 2q \exp(-\tilde{c}_6n^3\delta^2).$$

To bound $I_3$, note that

$$I_3 = r_k^T \hat{\Sigma}^{-1}_{R(Y)} (\Sigma_{R(Y)} - \hat{\Sigma}_{R(Y)}) \Sigma_{R(Y)}^{-1} r_k.$$

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

$$|I_3| \leq \|\hat{\Sigma}^{-1}_{R(Y)}\| \cdot \|\Sigma_{R(Y)} - \hat{\Sigma}_{R(Y)}\| \cdot \|\Sigma_{R(Y)}^{-1}\| \cdot \|r_k\|^2.$$

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

$$\Pr \left( \|\Sigma_{R(Y)} - \hat{\Sigma}_{R(Y)}\| \geq q\left(\frac{6}{n} + \delta\right) \right)$$

$$\leq q^2 \Pr \left( |r_s(Y_j, Y_l) - \hat{r}_s(Y_j, Y_l)| \geq \frac{6}{n} + \delta \right)$$

$$\leq 2q^2 \exp(-\tilde{c}_5n\delta^2) + 2q^2 \exp(-\tilde{c}_6n^3\delta^2).$$

Hence, the union bound of probability for $I_3$ is

$$\Pr \left( |I_3| \geq (c_8 + 1)c_0^{-2}q^4\left(\frac{6}{n} + \delta\right) \right) \leq 2q^2 \exp(-c_3nq^{-4}) + 2q^2 \exp(-c_4n^3q^{-4})$$

$$+ 2q^2 \exp(-\tilde{c}_5n\delta^2) + 2q^2 \exp(-\tilde{c}_6n^3\delta^2).$$
The final probability bound

$$\Pr \left( \left| \hat{\omega}_m^{\text{RCC1}} - \omega_m^{\text{RCC1}} \right| \geq c_{10} q^2 \left( \frac{6}{n} + \delta \right)^2 + 2c_{10} q^2 \left( \frac{6}{n} + \delta \right) + c_{11} q^4 \left( \frac{6}{n} + \delta \right) \right)$$

$$\leq 6 q^2 \left( \exp(-c_3 n q^{-4}) + \exp(-c_4 n^3 q^{-4}) \right) + (2 q^2 + 4 q) \left( \exp(-c_5 n \delta^2) + \exp(-c_6 n^3 \delta^2) \right),$$

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

$$\Pr \left( \left| \hat{\omega}_k^{\text{RCC1}} - \omega_k^{\text{RCC1}} \right| \geq c_2 q^4 n^{-\kappa} \right)$$

$$\leq 6 q^2 \left( \exp(-c_3 n q^{-4}) + \exp(-c_4 n^3 q^{-4}) \right) + (2 q^2 + 4 q) \left( \exp(-c_5 n^{1-2\kappa}) + \exp(-c_6 n^{3-2\kappa}) \right),$$

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

$$\Gamma_n = \left\{ \max_{k \in \{j, \ldots, p\}} \left| \hat{\omega}_k^{\text{RCC1}} - \omega_k^{\text{RCC1}} \right| \leq \frac{\delta A q^4 n^{-\kappa}}{2} \right\},$$

we have

$$\min_{k \in A} \hat{\omega}_k^{\text{RCC1}} \geq \min_{k \in A} \left\{ \omega_k^{\text{RCC1}} - \left| \hat{\omega}_k^{\text{RCC1}} - \omega_k^{\text{RCC1}} \right| \right\}$$

$$\geq \max_{k \in A^c} \omega_k^{\text{RCC1}} + \frac{\delta A q^4 n^{-\kappa}}{2} \geq \max_{k \in A^c} \hat{\omega}_k^{\text{RCC1}}.$$
Hence, there must exists \( \nu_n \geq t_n \), such that \( \tilde{A} = \hat{A}_{\nu_n} \). Moreover, for any \( t_n \leq \nu_n \), \( \hat{A}_{\nu_n} \subset \hat{A}_{t_n} \), which implies \( A \subset \tilde{A} \subset \hat{A}_{t_n} \). Therefore, let \( c_2 = \delta_A/2 \), by the choice of \( t_n = c_1 q^4 n^{-\kappa} \), \( c_1 \leq \delta_A/2 \), we have

\[
P(A \subset \hat{A}_{t_n}) \geq P(\Gamma_n)
\]

\[
\geq 1 - p \cdot \left\{ 6q^2 \left( \exp(-c_3 n q^{-4}) + \exp(-c_4 n^3 q^{-4}) \right) + (2q^2 + 4q) \left( \exp(-c_5 n^{1-2\kappa}) + \exp(-c_6 n^{3-2\kappa}) \right) \right\}.
\]

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

\[
\sum_{k=1}^{p} \omega_{mRCC1}^k \leq \sum_{k=1}^{p} \| \Sigma_{R(Y)}^{-1} \| \cdot \| r_k \|^2 \leq c_0^{-1} s q^2 = O(sq^2).
\]

This indicates that the number of \( \{ k : \omega_{mRCC1}^k > \epsilon q^4 n^{-\kappa} \} \) cannot exceed \( O(sq^{-2} n^\kappa) \) for any \( \epsilon > 0 \). Therefore, on the set

\[
\Delta_n = \left\{ \max_{1 \leq k \leq p} | \hat{\omega}_{mRCC1}^k - \omega_{mRCC1}^k | \leq \epsilon q^4 n^{-\kappa} \right\},
\]

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

\[
Pr ( | \tilde{A}_{t_n} | \leq O(sq^{-2} n^\kappa) ) \geq Pr(\Delta_n).
\]
The conclusion follows from the first part of Theorem 1.

References


Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimen-


REFERENCES


REFERENCES


