Comparison of Different Feaure Screening Methods with Application to BBS Associated Genes

XIUFANG WANG

Renmin University of China

April 30, 2022

1. Introduction

With rapid development of modern information technology, high-dimensional data have been popular in various scientific fields, especially in genomics. In high-dimensional data, the number of features p usually grows much faster than the sample size n, which has imposed great challenges on statistical inference. Therefore, in practice, how to extract useful information from the high-dimensional data becomes a critical step during data preprocessing.

In literature, a great number of feature screening procedures have been proposed, which can be roughly divided into two categories. The first category are model-based screening procedures. For example, with the linear model assumption, Fan and Lv (2008) proposed a sure independent screening procedure (SIS) based on the pearson correlation coefficient. Then, Li et al. (2012) improved the SIS procedure by replacing the pearson correlation coefficient with Kendall's rank correlation. Furthermore, Fan et al. (2011) and He et al. (2013) proposed nonparametric screening(NIS) procedures for additive models. However, these model-based methods are effective only when the designed model is close to the true one. Otherwise, the story changes drastically.

To reduce the risk of model misspecification, the second category procedures, which are model-free, have been developed. For instance, Zhu et al. (2011) proposed a sure independent feature ranking and screening approach(SIRS). Li et al. (2012) introduced distance correlation screening method(DC-SIS), which can be used for grouped predictor variables and multivariate response variables. Shao and Zhang (2014) developed martingale difference correlation based screening procedure (MDC-SIS). These methods are really preferred when we are lack of prior information of the regression structure. Nevertheless, most of them have a poor performance in the presence of outliers. To fill in the vulnerability, Zhou et al. (2020) introduced a robust meature called cumulative divergence (CD) to characterize mean dependence and proposed a CD-based forward screening procedure(CD-SIS). Moreover, Zheng et al. (2012) derived a generalized measure of correlation (GMC) for asymmetry, nonlinearity, and Beyond.

To have a good knowledge of these correlation measures, we conduct comprehensive simulations and apply these methods on a real data to identify their performance under different scenarios. In addition, we apply these procedures to investigate the associated genes with known causative genes for a rare disease Bardet-Biedl syndrome (BBS).

2. Data description

Bardet-Biedl syndrome (BBS) is a rare, inherited condition that can affect most organs in the body. So far, mutations in 21 genes have been identified as causing up to 80% of BBS cases. The main symptom of this disease is progressive visual impairment among people.

In the previous work, many researchers aim to find out additive genes having high association with the known causative genes based on the pairwise correlations. In this way, the corresponding gene therapy could be developed for the disease treatment. Here, we utilize the microarray expression data for an eQTL experiment in rat eye reported in Scheetz et al. (2006). For this dataset, 120 twelve-week old male rats were selected for tissue harvesting from the eyes and for microarray analysis. In the Affymetrix expression microarray, there are totally 31099 different gene probe sets from the mRNA of the eye tissues, of which the detailed information can be obtained from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5680. In literature, this dataset has been analyzed by several statistians, including but not limit to, Huang et al. (2010), Fan et al. (2011) and Shao and Zhang (2014).

In our project, we first compare the screening performance with different measures in terms of selecting related genes of TRIM 32(or BBS11) since some significant gene probes have been described in Fan et al. (2011). Then, we combine each screening method with group lasso algorithm and apply it to identify the potential genes having high association with BBS14, which causes up to 6% BBS disease.

3. Feature screening procedures

3.1. A brief review

Sure independence screening (SIS) was the original screening procedure based on the classical Pearson coefficient of correlation, which was proposed by Fan and Lv (2008) for linear models. Its idea is to order the covariates by their correlation coefficients and retain the higher ones with a given screen size d.

Then, to reduce the sensitivity to the outlies of the observations, Li et al. (2012) replaced the Pearson coefficient of correlation with robust rank correlation(τ), that is, robust rank correlation screening (RRCS). However, these two methods are both model-based and take risk of model misspecification.

To avoid model misspecification, several model-free methods have been proposed later, which could be classified as two categories according to their target set. The first class aims to filter out the active subset resulting from the conditional distribution independence. The second class focus on the covariates which contribute to the conditional mean of the response.

Next we first introduce the first class screening procedures, the representative one of which is the sure independence ranking and screening (SIRS) procedure proposed by Zhu et al. (2011). This method is robust to the outlies in the response observations since it only uses their rank. Assume $E(X_k) = 0$ and $\operatorname{var}(X_k) = 1$ for $k = 1, \ldots, p$. Define $\mathbf{\Omega}(y) = \mathbb{E}\{\mathbf{x}F(y \mid \mathbf{x})\} = \operatorname{cov}\{\mathbf{x}, \mathbf{1}(Y < y)\}$. The population quantity of its marginal utility measure for the predictors can be expressed as

$$\omega_k = \mathbb{E}\left\{\Omega_k^2(Y)\right\}, \quad k = 1, \ldots, p. \tag{1}$$

where $\Omega_k(y)$ is the k th element of $\mathbf{\Omega}(y)$.

And its sample version is as follows.

$$\tilde{\omega}_k = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n} X_{ik}\mathbf{1}\left(Y_i < Y_j\right)\right\}^2, \quad k = 1, \ldots, p$$

where X_{ik} denotes the k th element of $\mathbf{x}_i$.

Another representive measure of the first class is distance correlation (DC), a symmetric dependence measure, was introduced by Li et al. (2012). Assume that both $\mathbf{x}$ and $\mathbf{y}$ have finite first moments, the distance correlation (DC) is defined as

$$\operatorname{dcorr}(\mathbf{x}, \mathbf{y}) = \frac{\operatorname{dcov}(\mathbf{x}, \mathbf{y})}{\sqrt{\operatorname{dcov}(\mathbf{x}, \mathbf{y}) \operatorname{dcov}(\mathbf{y}, \mathbf{y})}} \tag{2}$$

where

$$\operatorname{dcov}^2(\mathbf{x}, \mathbf{y}) = \int_{R^{d_x+d_y}} \left\|\phi_{\mathbf{x},\mathbf{y}}(\mathbf{t}, \mathbf{s}) - \phi_{\mathbf{x}}(\mathbf{t})\phi_{\mathbf{y}}(\mathbf{s})\right\|^2 w(\mathbf{t}, \mathbf{s}) d\mathbf{t} d\mathbf{s} \tag{3}$$

In the above equation, d_x and d_y are the dimensions of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $w(\mathbf{t}, \mathbf{s}) = \left\{c_{d_x} c_{d_y} \|\mathbf{t}\|_{d_x}^{1+d_x} \|\mathbf{s}\|_{d_y}^{1+d_y}\right\}^{-1}$ with $c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$. $\|\mathbf{a}\|_d$ stands for the Euclidean norm of $\mathbf{a} \in \mathbb{R}^d$, and $\|\phi\|^2 = \phi\bar{\phi}$ for a complex-valued function ϕ with $\bar{\phi}$ being the conjugate of ϕ.

Recently, a new coefficient of correlation was proposed by Chatterjee (2021). It is a simple and interpretable measure of the degree of dependence between the variables, which can be defined as

$$\xi(X,Y) := \frac{\int \operatorname{var}\left(\mathbb{E}\left(1_{\{Y\geq t\}} \mid X\right)\right) d\mu(t)}{\int \operatorname{var}\left(1_{\{Y\geq t\}}\right) d\mu(t)} \tag{4}$$

where the data is rearranged as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ such that $X_{(1)} \leq \ldots \leq X_{(n)}$.

Besides, the corresponding sample version is given by

$$\xi_n(X,Y) := 1 - \frac{3\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1} \tag{5}$$

where r_i is the rank of $Y_{(i)}$.

Then, we introduce some screening methods focusing on the conditional mean function. Three methods are mainly considered here, including MDC-based SIS proposed by Shao and Zhang (2014), CD-based forward screening proposed by Zhou et al. (2020) and GMC-based screening with slicing estimation, which is newly-derived in this report.

Martingale distance correlation(MDC) is a natural extension of distance correlation, which is used to measure the departure of conditional mean independence between the response and the predictor. Define the martingale difference divergence of Y given X by

$$\operatorname{MDD}(Y \mid X)^2 = \frac{1}{c_q} \int_{\mathbf{R}^q} \frac{|g_{Y,X}(s) - g_Y g_X(s)|^2}{|s|_q^{1+q}} ds \tag{6}$$

where $g_{Y,X}(s) = \mathbb{E}\left(Ye^{i<s,X>}\right), g_Y = \mathbb{E}(Y)$, and $g_X(s) = \mathbb{E}\left(e^{i<s,X>}\right)$.

Then ,the martingale difference correlation of Y given X can be represented by

$$\mathrm{MDC}(Y \mid X)^2 = \begin{cases} \frac{\mathrm{MDD}(Y|X)^2}{\sqrt{\mathrm{var(Y)^2\,dvar(X)^2}}} & \text{if } \operatorname{var}(Y)^2 \operatorname{dvar}(X)^2 > 0, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Since the martingale difference correlation is defined under the assumption that both X and Y have finite second moments. It may lose efficiency in presence of outliers in the observations. Recently, culmulative divergence (CD) was developed by Zhou et al. (2020) to measure the departure from the relationship $\mathbb{E}(Y \mid \mathbf{x}) \stackrel{\text{a.s.}}{=} \mathbb{E}(Y)$. It is robust to the outliers of observations of predictors. Its definition is formularized by

$$\mathrm{CD}(Y \mid X) \stackrel{\text{def}}{=} \mathbb{E}\left[\operatorname{cov}^2\{Y, \mathbf{1}(X < \widetilde{X}) \mid \widetilde{X}\}\right] / \operatorname{var}(Y) \tag{8}$$

The estimator of $\mathrm{CD}(Y \mid X)$ is also given in Zhou et al. (2020) as follows.

$$\widehat{\mathrm{CD}}(Y \mid X) \stackrel{\text{def}}{=} n^{-3} \sum_{j=1}^{n} \left[\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right) \left\{I\left(X_i < X_j\right) - F_n\left(X_j\right)\right\}\right]^2 / \widehat{\operatorname{var}}(Y) \tag{9}$$

Finally, we introduce the generalized measure of correlation (GMC) proposed by Zheng et al. (2012).

In the regression model $Y = \mathbb{E}(Y \mid X) + \epsilon$ where $\mathbb{E}(\epsilon \mid X) = 0$, the $\mathrm{GMC}(Y \mid X)$ can be interpreted as the explained variance of Y by X. It can be formulized as

$$\mathrm{GMC}(Y \mid X) \stackrel{\text{def}}{=} 1 - \frac{\mathbb{E}\{Y - \mathbb{E}(Y \mid X)\}^2}{\operatorname{var}(Y)} \tag{10}$$

whenever $0 < \operatorname{var}(Y) < \infty$.

According to Zhu and Ng (1995), the slicing procedure proceeds as follows. First, we order the random sample $(X_i, Y_i), i = 1, \ldots, n$, by the values of X_is and denote the ordered data as $(X_{(i)}, Y_{(i)}), i = 1, \ldots, n,$, where $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ and $Y_{(i)}$ is the concomitant of $X_{(i)}$. Then, we divide the ordered data into H slices by the values of $X_{(i)}$s. We simply assume that $n = Hc$, so that there are c observations in each slice. Following the double subscripts in Zhu and Ng (1995), we can rewrite $X_{(h,j)} = X_{(c(h-1)+j)}$ and $Y_{(h,j)} = Y_{(c(h-1)+j)}$ for $h = 1, 2, \ldots, H$ and $j = 1, 2, \ldots, c$. Thus, the observations in the hth slice are $(X_{(h,j)}, Y_{(h,j)}), i = 1, 2, \ldots, c$. Taking average of the variance estimator in each slice, we can obtain the slicing estimator of Λ as

$$\Lambda_n = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{c-1}\sum_{j=1}^{c}(Y_{(h,j)} - \frac{1}{c}\sum_{j=1}^{c}Y_{(h,j)})^2 \tag{11}$$

Therefore, $\widehat{\mathrm{GMC}}(Y \mid X) = 1 - \Lambda_n / \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

3.2. A simple comparison of screening methods

To have a great knowledge of the difference among aforementioned screening methods, we show the result of the comparison in terms of five important properties in Table 1.

Below, We first define three active sets as follows.

$$\begin{aligned}
\mathcal{F}_1 &= \{1 \le i \le p : \beta_i \ne 0\} \\
\mathcal{F}_2 &= \{k : F(y \mid \mathbf{x}) \text{ functionally depends on } \mathbf{X_k} \text{ for some } \mathbf{y} \in \mathrm{supp}(\mathbf{Y})\} \\
\mathcal{F}_3 &= \{k : \mathbb{E}(y \mid \mathbf{x}) \text{ functionally depends on } \mathbf{X_k} \text{ for some } \mathbf{y} \in \mathrm{supp}(\mathbf{Y})\}
\end{aligned} \tag{12}$$

Table 1 reveals that there are significant differences considering their individual active set, which contains all important predictors as defined in (12). For instance, since both the SIS and RRCS procedures assume a linear model or a transformation regression model, their active sets contain the predictors whose regression coefficients

Table 1: Comparison of screening procedures.

Methods	**Model-free**	**Active set**	**Range**	**Robust of X**	**Tuning param**
SIS	F	$\mathcal{F}_1$	[-1,1]	F	F
RRCS	F	$\mathcal{F}_1$	[-1,1]	T	F
SIRS	T	$\mathcal{F}_2$	$[0, \infty]$	F	F
DC-SIS	T	$\mathcal{F}_2$	[0,1]	F	F
MDC-SIS	T	$\mathcal{F}_3$	[0,1]	F	F
CD-SIS	T	$\mathcal{F}_3$	[0,1/4]	T	F
GMC-SIS	T	$\mathcal{F}_3$	[0,1]	T	slice number
ξ_n-SIS	T	$\mathcal{F}_2$	[0,1]	T	F

are nonzero. Besides, the range of each marginal utility measure is stated in Table 1 and some of them are same. Furthermore, certain procedures are sensitive to the outliers in observations of predictors, such as SIS, DC-SIS and MDC-SIS. However, the extreme values are frequent in high-dimensional situation. Thus, in practice, we may recommond the robust screening procedures for a higher efficiency. Meanwhile, we notice that the GMC-based marginal screening method induces an annoying tuning parameter, that is, the number of observations within each slice c. Through several simulations, we find that the accuracy of our procedure increases as c becomes larger for a wide range. Also, this parameter has little influence on the consistency and asymptotic normality of the slicing estimator of GMC. Therefore, it would not increase too much computation cost since a reasonable range of c is given in Zhu and Ng (1995).

4. Real data analysis

Example 1. *(BBS associated genes)*

To compare the power of the proposed marginal sure independence screening procedures in high-dimensional setting, we first follow the conclusion in Fan et al. (2011), which identified nine significant gene probes related to BBS11 shown in Table 2 . We set the screening size for genes as $d = \lfloor n/\log(n) \rfloor$ and $d = n$ respectively. The results in Table 2 illustrate that the GMC*-based screening procedure could select eight out of* *nine genes, which is comparable with* MDC-SIS *and* DC-SIS *when* $d = n$*. Besides, if* $d = \lfloor n/\log(n) \rfloor$*, then the performance of the* GMC-SIS *is superior than other methods, including* MDC-SIS *with longer running time.*

Table 2: Results of gene screening using seven methods for Example 1

Gene Probe ID	$d = \lfloor n/\log(n) \rfloor$							
	SIS	RRCS	SIRS	DC-SIS	MDC-SIS	CD-SIS	GMC-SIS	ξ_n-SIS
1371755_at							$\surd$	
1372928_at								
1373534_at		$\surd$		$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
1373944_at								
1374669_at							$\surd$	
1376686_at	$\surd$			$\surd$	$\surd$	$\surd$		
1376747_at	$\surd$			$\surd$	$\surd$			
1377880_at					$\surd$		$\surd$	
1378590_at	$\surd$	$\surd$		$\surd$	$\surd$	$\surd$	$\surd$	
Gene Probe ID	$d = n$							
	SIS	RRCS	SIRS	DC-SIS	MDC-SIS	CD-SIS	GMC-SIS	ξ_n-SIS
1371755_at							$\surd$	
1372928_at		$\surd$		$\surd$	$\surd$	$\surd$	$\surd$	
1373534_at	$\surd$	$\surd$		$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
1373944_at		$\surd$		$\surd$	$\surd$	$\surd$	$\surd$	
1374669_at		$\surd$		$\surd$	$\surd$		$\surd$	$\surd$
1376686_at	$\surd$	$\surd$		$\surd$	$\surd$	$\surd$	$\surd$	
1376747_at	$\surd$	$\surd$		$\surd$	$\surd$	$\surd$		
1377880_at	$\surd$	$\surd$		$\surd$	$\surd$	$\surd$	$\surd$	
1378590_at	$\surd$	$\surd$		$\surd$	$\surd$	$\surd$	$\surd$	$\surd$

Since the target sets of some procedures are little different, we proceed to evaluate their performance in predicting the expression level of gene BBS14. In this way, additive causative genes may be found by the pairwise correlation with BBS14, which will give some inspiration for the BBS gene therapy.

Specifically, we adpot the group lasso algorithm for the retained data from feature screening step with screen size $d = 500, n, \lfloor n/\log(n) \rfloor$ *respectively. For our* GMC-SIS *method, here we consider* $c = 12$ *and* $c = 20$ *two cases. Five-fold cross-validation was used to select a penalty parameter for group lasso algorithm.*

We report the average model size (Ave.S) and mean square prediction error (Ave.MSPE) out of 200 replications in Table 3. For simplicity, we denote each screening method with their abbreviations and the group lasso algorithm as GL.

Table 3: Average model size(Ave.S) and average mean square prediction error(Ave.MSPE) with $d = \lfloor n/\log(n) \rfloor$, n and 500 out of 200 replications for Example 1

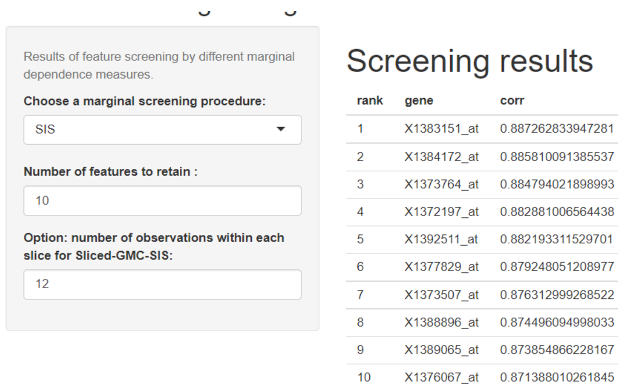
.

Methods	$d = 500$		$d = n$		$d = \lfloor n/\log(n) \rfloor$	
	Ave.S	Ave.MSPE	Ave.S	Ave.MSPE	Ave.S	Ave.MSPE
SIS-GL	30.98	0.232	22.65	0.245	13.01	0.207
RRCS-GL	31.5	0.231	22.48	0.256	13.42	0.261
SIRS-GL	45.58	0.257	27.02	0.239	12.2	0.284
DC-GL	45.52	0.244	13.92	0.109	5.77	0.168
MDC-GL	27.37	0.246	25.66	0.302	8.62	0.309
CD-GL	46.68	0.247	31.74	0.238	12.38	0.372
GMC-GL(c=12)	29.41	0.239	24.15	0.248	13.09	0.268
GMC-GL(c=20)	31.22	0.232	13.98	0.253	11.9	0.253
ξ_n-GL	33.08	0.241	22.58	0.256	12.36	0.230

From Table 3, it seems that for most screening procedures, the average prediction error increases and average model size decreases as the original screen size d decreases. Specifically, the GMC-based screening combined with group lasso procedure is competitive among all model-free method with a smaller average prediction error. Moreover, if we set the screen size d = n, then DC-GL method is the absolute winner. The others are comparable in terms of the average model size and prediction error. Furthermore, when we continue to shrink d, DC-GL method still has a great performance, followed by SIS-GL and ξ_n-GL methods.

Another direct result we could obtained is to compare the selected genes by each methods. For an intuitive exhibition of these different marginal screening procedures on the dataset, we construct a website based on shiny package in R. Here some key screenshots are shown as follows. For details, see https://github.com/wangxiufang123/Report2022.

5. Discussion and future work



Results of feature screening by different marginal dependence measures.

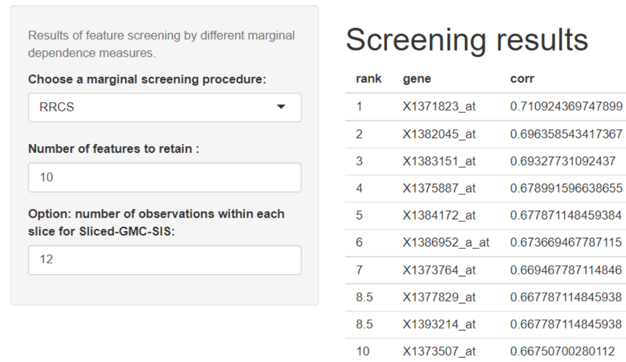
Choose a marginal screening procedure: SIS

Number of features to retain : 10

Option: number of observations within each slice for Sliced-GMC-SIS: 12

Screening results

rank	gene	corr
1	X1383151_at	0.887262833947281
2	X1384172_at	0.885810091385537
3	X1373764_at	0.884794021898993
4	X1372197_at	0.882881006564438
5	X1392511_at	0.882193311529701
6	X1377829_at	0.879248051208977
7	X1373507_at	0.876312999268522
8	X1388896_at	0.874496094998033
9	X1389065_at	0.873854866228167
10	X1376067_at	0.871388010261845



Results of feature screening by different marginal dependence measures.

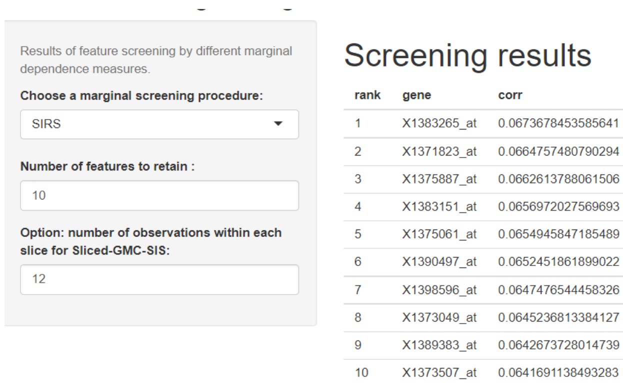
Choose a marginal screening procedure: RRCS

Number of features to retain : 10

Option: number of observations within each slice for Sliced-GMC-SIS: 12

Screening results

rank	gene	corr
1	X1371823_at	0.710924369747899
2	X1382045_at	0.696358543417367
3	X1383151_at	0.69327731092437
4	X1375887_at	0.678991596638655
5	X1384172_at	0.677871148459384
6	X1386952_a_at	0.673669467787115
7	X1373764_at	0.669467787114846
8.5	X1377829_at	0.667787114845938
8.5	X1393214_at	0.667787114845938
10	X1373507_at	0.66750700280112



Results of feature screening by different marginal dependence measures.

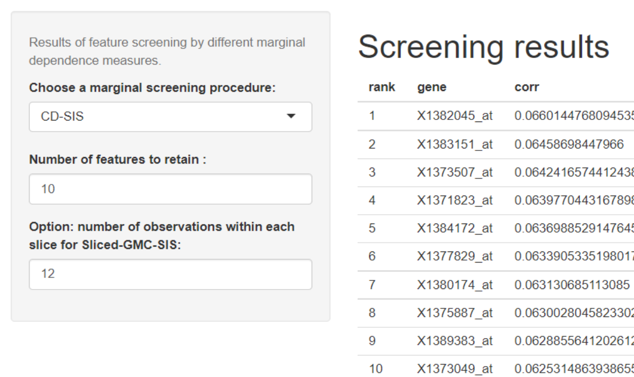
Choose a marginal screening procedure: SIRS

Number of features to retain : 10

Option: number of observations within each slice for Sliced-GMC-SIS: 12

Screening results

rank	gene	corr
1	X1383265_at	0.0673678453585641
2	X1371823_at	0.0664757480790294
3	X1375887_at	0.0662613788061506
4	X1383151_at	0.0656972027569693
5	X1375061_at	0.0654945847185489
6	X1390497_at	0.0652451861899022
7	X1398596_at	0.0647476544458326
8	X1373049_at	0.0645236813384127
9	X1389383_at	0.0642673728014739
10	X1373507_at	0.0641691138493283



Results of feature screening by different marginal dependence measures.

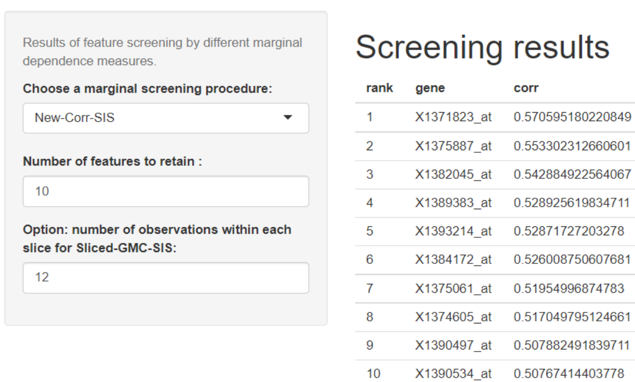
Choose a marginal screening procedure: CD-SIS

Number of features to retain : 10

Option: number of observations within each slice for Sliced-GMC-SIS: 12

Screening results

rank	gene	corr
1	X1382045_at	0.0660144768094535
2	X1383151_at	0.06458698447966
3	X1373507_at	0.0642416574412438
4	X1371823_at	0.0639770443167898
5	X1384172_at	0.0636988529147645
6	X1377829_at	0.0633905335198017
7	X1380174_at	0.063130685113085
8	X1375887_at	0.0630028045823302
9	X1389383_at	0.0628855641202612
10	X1373049_at	0.0625314863938655



Results of feature screening by different marginal dependence measures.

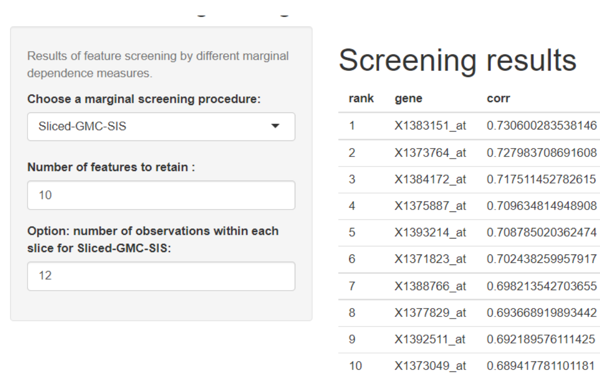
Choose a marginal screening procedure: New-Corr-SIS

Number of features to retain : 10

Option: number of observations within each slice for Sliced-GMC-SIS: 12

Screening results

rank	gene	corr
1	X1371823_at	0.570595180220849
2	X1375887_at	0.553302312660601
3	X1382045_at	0.542884922564067
4	X1389383_at	0.528925619834711
5	X1393214_at	0.52871727203278
6	X1384172_at	0.526008750607681
7	X1375061_at	0.51954996874783
8	X1374605_at	0.517049795124661
9	X1390497_at	0.507882491839711
10	X1390534_at	0.50767414403778



Results of feature screening by different marginal dependence measures.

Choose a marginal screening procedure: Sliced-GMC-SIS

Number of features to retain : 10

Option: number of observations within each slice for Sliced-GMC-SIS: 12

Screening results

rank	gene	corr
1	X1383151_at	0.730600283538146
2	X1373764_at	0.727983708691608
3	X1384172_at	0.717511452782615
4	X1375887_at	0.709634814948908
5	X1393214_at	0.708785020362474
6	X1371823_at	0.702438259957917
7	X1388766_at	0.698213542703655
8	X1377829_at	0.693668919893442
9	X1392511_at	0.692189576111425
10	X1373049_at	0.689417781101181

Combining DC-SIS and GMC-SIS with group lasso seem to have a superior performance among all methods when $d = n$ or $d = \lfloor n/\log(n) \rfloor$. Superisingly, since the parameter c in GMC-SIS could be predetermined, we expect this method to have a better performance by increasing c properly. Besides, Zhu and Ng (1995) also stated the effect of c on the convergence rate of the slicing estimator $\widehat{\mathrm{GMC}}$.

In literature, a great bulk of iterative versions of marginal screening procedures have been developed to address the three issues. First, some unimportant predictors *that are highly correlated with the important predictors can have higher priority for being selected by SIS than other important predictors that are relatively weakly related to the response. Second, an important predictor that is marginally uncorrelated but jointly correlated with the response cannot be picked by SIS and thus will not enter the estimated model. Third, the issue of collinearity between predictors adds difficulty to the problem of variable selection. Therefore, we consider to derive an iterative version for the GMC-based screening procedure and compare it with the counterpart of other methods.*

Meanwhile, since we only consider the group lasso algorithm in the second variable selection stage, which may lack accuracy of comparison, we should try other variable selection methods in the future. In addition, we could apply these procedures on more examples because of the heterogeneity of high-dimensional data.

REFERENCE

Chatterjee, S. (2021). A new coefficient of correlation. Journal of the American Statistical Association 116*(536), 2009–2022.*

Fan, J., Y. Feng, and R. Song (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. Journal of the American Statistical Association 106*(494), 544–557.*

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

He, X., L. Wang, and H. G. Hong (2013). Quantile-adaptive model-free variable *screening for high-dimensional heterogeneous data.* The Annals of Statistics 41*(1), 342–369.*

Huang, J., J. L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. Annals of statistics 38*(4), 2282.*

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

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

Scheetz, T. E., K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant, et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences 103*(39), 14429–14434.*

Shao, X. and J. Zhang (2014). Martingale difference correlation and its use in high-dimensional variable screening. Journal of the American Statistical Association 109*(507), 1302–1318.*

Zheng, S., N.-Z. Shi, and Z. Zhang (2012). Generalized measures of correlation for asymmetry, nonlinearity, and beyond. Journal of the American Statistical Association 107*(499), 1239–1252.*

Zhou, T., L. Zhu, C. Xu, and R. Li (2020). Model-free forward screening via cumulative divergence. Journal of the American Statistical Association 115*(531), 1393–1405.*

Zhu, L.-P., L. Li, R. Li, and L.-X. Zhu (2011). Model-free feature screening for ultrahigh-dimensional data. Journal of the American Statistical Association 106*(496), 1464–1475.*

Zhu, L.-X. and K. W. Ng (1995). Asymptotics of sliced inverse regression. Statistica Sinica, *727–736.*