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

Xiufang Wang

Institute of Statistics and Big Data

Renmin University of China

April 26, 2022

Presentation Overview

Introduction

1. Data background: 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.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.
2. Data source: from an eQTL experiment in rat eye loaded in <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5680>.
3. Data description:
   - sample size: 120 twelve-week old male rats
   - dimensionality: 31099 different gene probe sets

Introduction

4. Our target:
   - compare the screening performance with different marginal measures in terms of selecting related genes of TRIM 32(or BBS11)
   - 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.

Screening procedures with Marginal Utility Measure

Feature screening procedures can be roughly divided into two categories. The first is model-based.

1. Sure independence screening (SIS):
   - based on Pearson correlation coefficient
   - for linear models
   - $\mathbb{E}(X_k^2) < \infty$ for $k = 1, 2, \ldots, p$
   - sensitive to outlies or extreme values in observations
2. Robust rank correlation screening (RRCS):
   - based on Kendall τ correlation coefficient
   - for semiparametric models and single-index models

cons: model misspecification

Screening procedures with marginal utility measure

The second is model-free.

1. Sure independence ranking and screening (SIRS)
   - $\boldsymbol{\Omega}(y) = \mathbb{E}\{\mathbf{x}F(y \mid \mathbf{x})\} = \text{cov}\{\mathbf{x}, \mathbf{1}(Y < y)\}$.
   - population quantity:

$$\omega_k = \mathbb{E}\{\Omega_k^2(Y)\}, k = 1, \ldots, p$$

     where $\Omega_k(y)$ is the k th element of $\boldsymbol{\Omega}(y)$.
   - sample version:

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

     where X_{ik} denotes the k th element of $\mathbf{x}_i$.
   - robust to the outlies in the response observations.

Screening procedures with marginal utility measure

2. Distance correlation based SIS (DC-SIS)
   - conditions:X and Y have finite first moments.
   - based on distance correlation $\mathcal{R}(X, Y)$:

$$\mathcal{R}^2(X, Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0 \end{cases}$$

   where

$$\mathcal{V}^2(X, Y) = \|f_{X,Y}(t, s) - f_X(t)f_Y(s)\|^2$$

   - $f_X(\cdot)$, $f_Y(\cdot)$ and $f_{X,Y}(\cdot, \cdot)$ are characteristic functions of X, Y and (X, Y).
   - for multivariate responses; screening grouped variables

Screening procedures with marginal utility measure

3. Martingale distance correlation based SIS (MDC-SIS)
   - conditions: $\mathbb{E}(|Y|^2 + |X|^2) < \infty$
   - based on martingale distance correlation $\mathrm{MDC}(Y \mid X)$:

$$[-\mathbb{E}\{Y - \mathbb{E}(Y)\}\{\tilde{Y} - \mathbb{E}(Y)\}\|X - \tilde{X}\|]^{1/2} / \{\mathrm{var}(Y)\,\mathrm{dvar}(X)\}^{1/2}$$

   - sample version: $\widehat{\omega}_k = MDC_n(Y \mid X_k)^2$

Screening procedures with marginal utility measure

4. Culmulative divergence based SIS (CD-SIS)
   - conditions: $\mathrm{var}(X) > 0$ and $0 < \mathrm{var}(Y) < \infty$
   - based on culmulative divergence $\mathrm{CD}(Y \mid X)$:

$$\mathbb{E}\left[\mathrm{cov}^2\{Y, \mathbf{1}(X < \widetilde{X}) \mid \widetilde{X}\}\right] / \mathrm{var}(Y)$$

   - sample version:

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

Screening procedures with marginal utility measure

5. Generalize measure of correlation based SIS (GMC-SIS)
   - conditions: $\mathrm{var}(X) > 0$ and $0 < \mathrm{var}(Y) < \infty$
   - based on generalize measure of correlation $\mathrm{GMC}(Y \mid X)$:

$$1 - \frac{\mathbb{E}\{Y - \mathbb{E}(Y \mid X)\}^2}{\mathrm{var}(Y)}$$

   - slicing estimator:

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

   where

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

   - tuning parameter: c or equivalently H; the larger c is, the greater the performance of the GMC-SIS is.

Screening procedures with marginal utility measure

6. A new correlation based SIS (ξ_n-SIS)
   - first rearrange the data as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ such that $X_{(1)} \leq \ldots \leq X_{(n)}$.
   - based on

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

   - sample version:

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

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

Comparison of screening procedures

$\mathcal{F}_1 = \{1 \leq i \leq p : \beta_i \neq 0\}$
$\mathcal{F}_2 = \{k : F(y \mid \mathbf{x})$ functionally depends on X_k for some $y \in \text{supp}(Y)\}$
$\mathcal{F}_3 = \{k : \mathbb{E}(y \mid \mathbf{x})$ functionally depends on X_k for some $y \in \text{supp}(Y)\}$

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

Table: Comparison of screening procedures

Comparison of screening procedures

Methods	**Robust of X**	**Invariance**	**Tuning param**
SIS	F	scale change of X and Y	F
RRCS	T	monotone trans of X and Y	F
SIRS	F	monotone trans of Y	F
DC-SIS	F	scale change of X and Y	F
MDC-SIS	F	scale change of X and Y	F
CD-SIS	T	monotone trans of X	F
GMC-SIS	T	monotone trans of X	slice number
ξ_n-SIS	T	monotone trans of X and Y	F

Table: Comparison of screening procedures

Results

Nine genes have been selected as the most relevant ones of gene BBS11 by INIS-penGAM algorithm in Fan, Feng, and Song (2011).

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							√	
1372928_at								
1373534_at		√		√	√	√	√	√
1373944_at								
1374669_at							√	
1376686_at	√			√	√	√		
1376747_at	√			√	√			
1377880_at					√		√	
1378590_at	√	√		√	√	√	√	
Gene Probe ID	$d = n$							
	SIS	RRCS	SIRS	DC-SIS	MDC-SIS	CD-SIS	GMC-SIS	ξ_n-SIS
1371755_at							√	
1372928_at		√		√	√	√	√	
1373534_at	√	√		√	√	√	√	√
1373944_at		√		√	√	√	√	
1374669_at		√		√	√		√	√
1376686_at	√	√		√	√	√	√	
1376747_at	√	√		√	√	√		
1377880_at	√	√		√	√	√	√	
1378590_at	√	√		√	√	√	√	√

Figure: Results of gene screening using different methods

Results

Combine each feature screening method with group lasso algorithm.

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

Figure: 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

.

Results

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

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

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

Discussion and future work

- 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. [Zhu, 1995] stated the effect of c on the convergence rate of the slicing estimator $\widehat{\text{GMC}}$.
- derive an iterative version for the GMC-based screening procedure and compare it with the counterpart of other methods.
- adopt other variable selection methods in the second stage.
- apply these procedures on more examples.

References I


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


Li, Gaorong and Peng, Heng and Zhang, Jun and Zhu, Lixing (2012)
Robust rank correlation based screening.
The Annals of Statistics 40(3),1846–1877.


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


Zhu, Li-Ping and Li, Lexin and Li, Runze and Zhu, Li-Xing (2011)
Model-free feature screening for ultrahigh-dimensional data.
Journal of the American Statistical Association 106(496),1464–1475.


Li, Runze and Zhong, Wei and Zhu, Liping (2012)
Feature screening via distance correlation learning.
Journal of the American Statistical Association 107(499),1129–1139.

References II


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


Zhou, Tingyou and Zhu, Liping and Xu, Chen and Li, Runze (2020)
Model-free forward screening via cumulative divergence.
Journal of the American Statistical Association 115(531),1393–1405.


Zheng, Shurong and Shi, Ning-Zhong and Zhang, Zhengjun (2012)
Generalized measures of correlation for asymmetry, nonlinearity, and beyond.
Journal of the American Statistical Association 107(499),1239–1252.


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


Zhu, Li-Xing and Ng, Kai W (1995)
Asymptotics of sliced inverse regression.
Statistica Sinica ,727–736.

Thanks

Questions? Comments?