Sure independence screening for ultrahigh dimensional feature space

Wenbin Zhou PB19151764
Jiazhi He PB19151772
Binghe Zhu PB19051193

USTC

April 5, 2022

Overview

Overview

Background

Consider the problem of estimating a p -vector of parameters $\boldsymbol{\beta}$ from the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y} = (Y_1, \cdots, Y_n)^T$ is an n-vector of responses, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ is an $n \times p$ random design matrix with IID $\mathbf{x}_1, \cdots, \mathbf{x}_n$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-vector of parameters and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is an n-vector of IID random errors.

- When dimension p is high, it is often assumed that only a small number of predictors among $X_1, \cdots, X_p$ contribute to the response, which amounts to assuming ideally that the parameter vector $\boldsymbol{\beta}$ is sparse.
- With sparsity,variable selection can improve the accuracy of estimation by effectively identifying the subset of important predictors, and also enhance model interpretability with parsimonious representation.

Background

Appealing features of the Dantzig selector include that

- it is easy to implement because the convex optimization that the Dantzig selector solves can easily be recast as a linear program and
- it has the oracle property in the sense of Donoho and Johnstone(1994).

We still have four concerns when the Dantzig selector is applied to high or ultrahigh dimensional problems.

1. a potential hurdle is the computation cost for large or huge scale problems such as implementing linear programs in dimension of tens of hundreds of thousands.
2. the factor $\log(p)$ can become large and may not be negligible when dimension p grows rapidly with sample size n.
3. As dimensionality grows, their uniform uncertainty principle condition may be difficult to satisfy.
4. There is no guarantee that the Dantzig selector picks up the right model though it has the oracle property.

Dimensionality reduction

Dimension reduction or feature selection is an effective strategy to deal with high dimensionality. With dimensionality reduced from high to low, the computational burden can be reduced drastically. Meanwhile, accurate estimation can be obtained by using some well-developed lower dimensional method. Motivated by this along with those concerns on the Dantzig selector, we have the following main goal in our paper:

Main goal

reduce dimensionality p from a large or huge scale(say, $\exp\{O(n^{\xi})\}$ for some $\xi > 0$ to a relatively large scale d(e.g. $o(n)$) by a fast and efficient method.

Correlation learning

We achieve the goal by introducing the concept of sure screening and proposing a sure screening method.

Definition of SIS

It is based on correlation learning which filters out the features that have weak correlation with the response.Such correlation learning is called sure independence screening(SIS)

Here and below, by sure screening we mean a property that all the important variables survive after variable screening with probability tending to 1.

Overview

corralation learning

We introduce a simple sure screening method using componentwise regression or equivalently correlation learning.

- We centre each input variable so that the observed mean is 0, and we scale each predictor so that the sample standard deviation is 1.
- Let $\mathcal{M}_* = \{1 \leq i \leq p : \beta_i \neq 0\}$ be the true sparse model with non-sparsity size $s = |\mathcal{M}_*|$. The other $p - s$ variables can also be correlated with the response variable via linkage to the predictors that are contained in the model.
- Let $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_p)^T$ be a p-vector that is obtained by componentwise regression,i.e.

$$\boldsymbol{\omega} = \mathbf{X}^T \mathbf{y} \tag{2}$$

where the $n \times p$ data matrix $\mathbf{X}$ is first standardized columnwise. Hence, $\boldsymbol{\omega}$ is really a vector of marginal correlations of predictors with the response variable, rescaled by the standard deviation of the response.

Componentwise regression

Connections between compoentwise regression and marginal correlation: regression $\mathbf{y}$ on $\mathbf{x}_i$ to get the coefficient $\widetilde{\beta}_i$

$$\begin{aligned}
\widetilde{\beta}_i &= (\mathbf{x}_i'\mathbf{x}_i)^{-1}\mathbf{x}_i'\mathbf{y} \\
&= \left(\sum \mathbf{x}_i^2\right)^{-1} \mathbf{x}_i\mathbf{y} \\
&= \left(\frac{1}{n-1}\sum(\mathbf{x}_i - \overline{\mathbf{x}})^2\right)^{-1} \frac{1}{n-1}\mathbf{x}_i'\mathbf{y} \\
&= \frac{1}{n-1}\mathbf{x}_i'\mathbf{y} = \frac{1}{n-1}\mathbf{w}_i
\end{aligned}$$

where $\overline{\mathbf{x}} = 0$, $\frac{1}{n-1}\sum(\mathbf{x}_i - \overline{\mathbf{x}})^2 = 1$

correlation learning

For any given $\gamma \in (0, 1)$, we sort the p componentwise magnitudes of the vector $\boldsymbol{\omega}$ in a decreasing order and define a submodel

$$\mathcal{M}_\gamma = \{1 \leq i \leq p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\} \quad (3)$$

where $[\gamma n]$ denotes the integer part of γn.

- This is a straightforward way to shrink the full model $\{1, \cdots, p\}$ down to a submodel $\mathcal{M}_\gamma$ with size $d = [\gamma n] < n$.
- Such correlation learning ranks the importance of features according to their marginal correlation with the response variable and filters out those that have weak marginal correlations with the response variable. We call this correlation screening method SIS, since each feature is used independently as a predictor to decide how useful it is for predicting the response variable.

correlation learning

- The computational cost of SIS or correlation learning is that of multiplying a $p \times n$ matrix by an n-vector plus obtaining the largest d components of a p-vector, so SIS has computational complexity $O(pn)$.
- It is worth mentioning that SIS uses only the order of componentwise magnitudes of $\boldsymbol{\omega}$, so it is indeed invariant under scaling.
- To implement SIS, we note that linear models with more than n parameters are not identifiable with only n data points. Hence, we may choose $d = [\gamma n]$ to be conservative, for instance, $n - 1$ or $n/\log(n)$ depending on the order of sample size n.
- Although SIS is proposed to reduce dimensionality p from high to below sample size n, nothing can stop us applying it with final model size $d \geq n$, say $\gamma \geq 1$. It is obvious that larger d means larger probability of including the true model $\mathcal{M}_*$ in the final model $\mathcal{M}$.

Rationale of correlation learning

To understand better the rationale of correlation learning, we now introduce an iteratively thresholded ridge regression screener(ITRRS), which is an extension of the dimensionality reduction method SIS.

- When there are more predictors than observations, it is well known that the least squares estimator $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T\mathbf{y}$ is noisy, where $(\mathbf{X}^T\mathbf{X})^+$ denotes the Moore-Penrose generalized inverse of $\mathbf{X}^T\mathbf{X}$.
- We therefore consider ridge regression, namely linear regression with l_2-regularization to reduce the variance. Let $\boldsymbol{\omega}^\lambda = (\omega_1^\lambda, \cdots, \omega_p^\lambda)^T$ be a p-vector that is obtained by ridge regression,i.e.

$$\boldsymbol{\omega}^\lambda = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbf{y} \tag{4}$$

Rationale of correlation learning

It is obvious that

$$\boldsymbol{\omega}^{\lambda} \to \hat{\boldsymbol{\beta}}_{\mathrm{LS}} \quad \text{as } \lambda \to 0 \tag{5}$$

and the scaled ridge regression estimator tends to the componentwise regression estimator:

$$\lambda\boldsymbol{\omega}^{\lambda} \to \boldsymbol{\omega} \quad \text{as } \lambda \to \infty \tag{6}$$

- In view of property (5) to make $\boldsymbol{\omega}^{\lambda}$ less noisy we should choose a large regularization parameter λ to reduce the variance in the estimation.
- Note that the ranking of the absolute components of $\boldsymbol{\omega}^{\lambda}$ is the same as that of $\lambda\boldsymbol{\omega}^{\lambda}$. In light of property (6) the componentwise regression estimator is a specific case of ridge regression with $\lambda = \infty$, namely, it makes the resulting estimator as little noisy as possilbe.

Rationale of correlation learning

For any given $\delta \in (0,1)$, we sort the p componentwise magnitudes of the vector $\boldsymbol{\omega}^{\lambda}$ in a descending order and define a submodel

$$\mathcal{M}_{\delta,\lambda}^{1} = \{1 \leq i \leq p : \left|\omega_i^{\lambda}\right| \text{ is among the first } [\delta p] \text{ largest of all}\} \qquad (7)$$

The ITRRS is as follows

1. First, carry out the procedure in submodel (7) to the full model $\{1, \cdots, p\}$ and obtain a submodel $\mathcal{M}_{\delta,\lambda}^{1}$ with size $[\delta p]$.
2. Then, apply a similar procedure to the model $\mathcal{M}_{\delta,\lambda}^{1}$ and again obtain a submodel $\mathcal{M}_{\delta,\lambda}^{2} \subset \mathcal{M}_{\delta,\lambda}^{1}$ with size $[\delta^2 p]$, and so on.
3. Finally,obtain a submodel $\mathcal{M}_{\delta,\lambda} = \mathcal{M}_{\delta,\lambda}^{k}$ with size $d = [\delta^k p] < n$, where $[\delta^{k-1} p] \geq n$.

ITRRS

- The ITRRS provides a very nice technical tool for understanding how fast the dimension p can grow comparing with sample size n and how the final model size d can be chosen while the sure screening property still holds for correlation learning. The question of whether the ITRRS has the sure screening property as well as how the tuning parameters γ and δ should be chosen will be answerd by theorem 3.
- The number of steps in the ITRRS depends on the choice of $\delta \in (0, 1)$. We shall see in theorem 3 that δ cannot be chosen too small, which means that there shold not be too many iteration steps in the ITRRS. This is due to the cumulation of the probability errors of missing some important variables over the iterations.

Overview

PLS and SCAD

The original problem is based on the now much smaller submodel $\mathcal{M}$, namely

$$\mathbf{y} = \mathbf{X}_{\mathcal{M}}\beta + \varepsilon \tag{8}$$

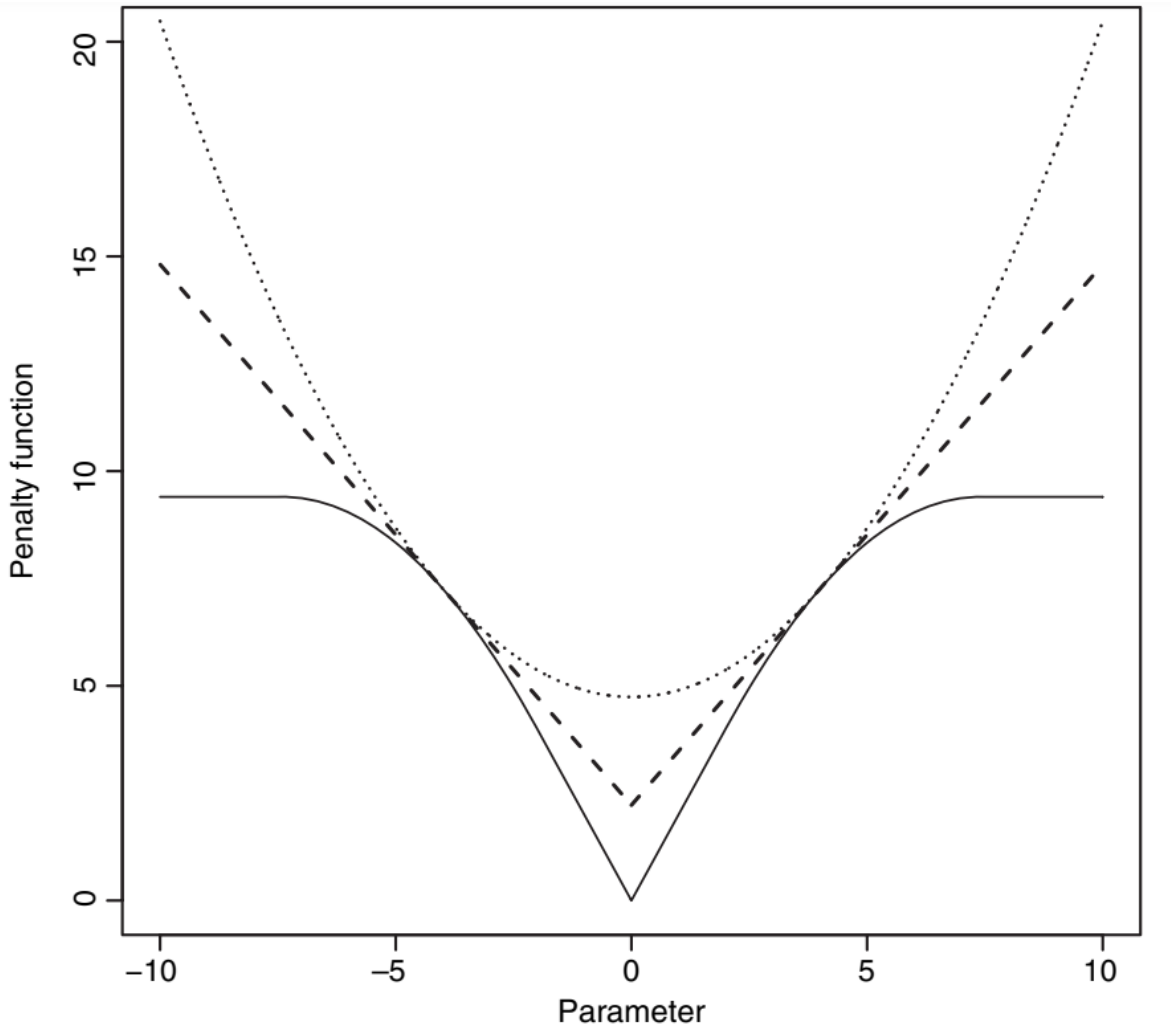
PLS problem

$$l(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\beta)^2 + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|) \tag{9}$$

problem (9) can be cast as a sequence of penalized L_1 regression problems

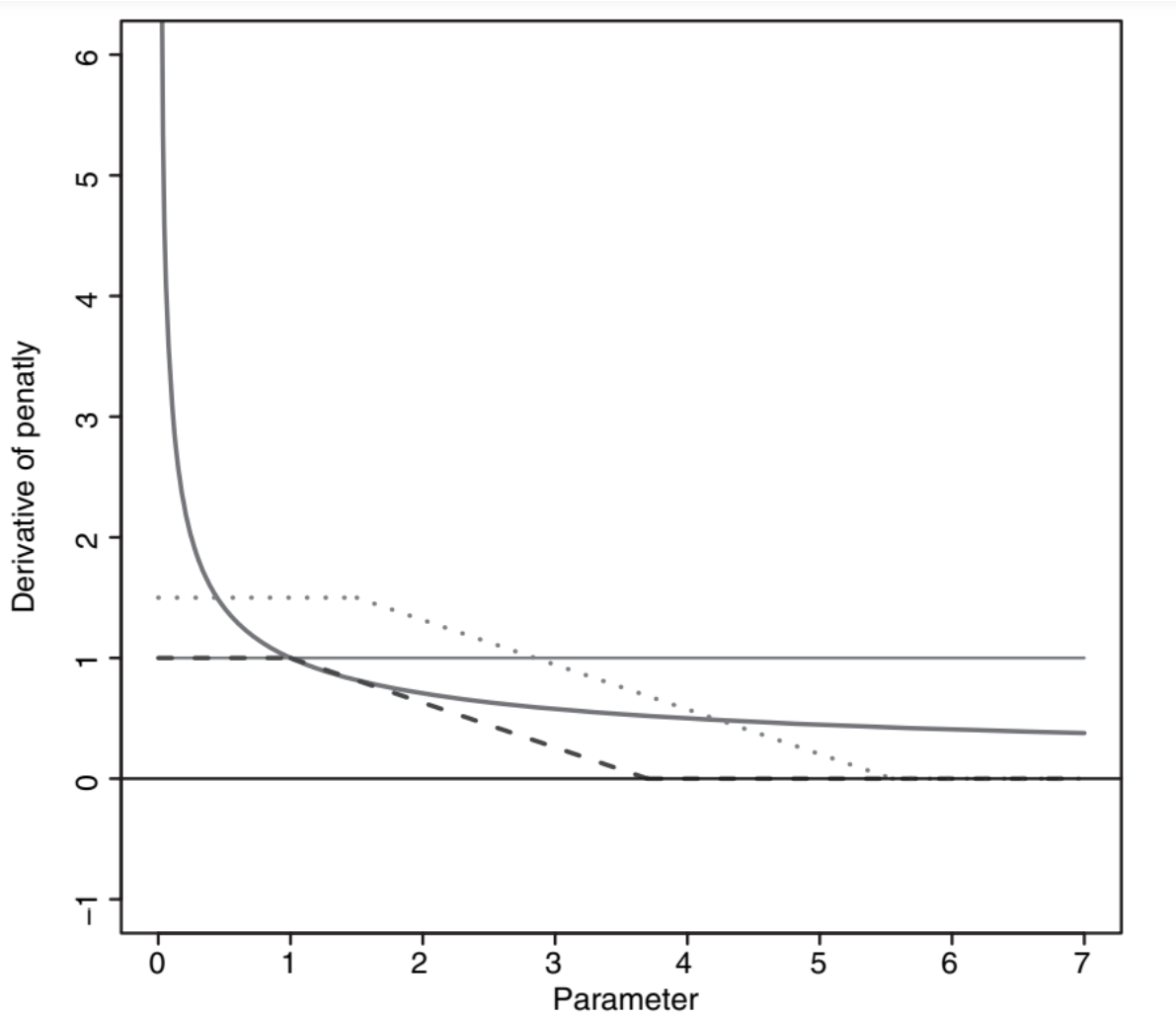
$$\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\beta)^2 + \sum_{j=1}^{d} w_j^{(k)}\,|\beta_j| \tag{10}$$

PLS and SCAD



SCAD penalty (———) and its local linear (-------) and quadratic (·····) approximations

PLS and SCAD



(b) $p'_\lambda(\cdot)$ for penalized L_1 (———), SCAD with $\lambda = 1$ (-------) and $\lambda = 1.5$ (·····) and adaptive lasso (———) with $\gamma = 0.5$

PLS and SCAD

Fan (1997) proposed a continuously differentiable penalty function called the SCAD penalty, which is defined by

$$p'_\lambda(|\beta|) = \lambda \left\{ I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right\} \quad \text{for some } a > 2 \tag{11}$$

The minimum concave penalty in Zhang (2007) is given by

$$p'_\lambda(|\beta|) = (a\lambda - |\beta|)_+ / a$$

Adaptive lasso and Dantzig selector

The adaptively weighted l_1-penalty used in the PLS problem (9)

$$\lambda \sum_{j=1}^{d} \omega_j |\beta_j|$$

$\hat{\beta}_{\mathrm{DS}}$ is the solution to the l_1 regularization problem.

$$\min_{\boldsymbol{\zeta} \in \mathbb{R}^d} (\|\boldsymbol{\zeta}\|_1) \quad \text{subject to } \|\mathbf{X}^T \mathbf{r}\|_\infty \leq \lambda_d \sigma \tag{12}$$

The above convex optimization problem can easily be recast as a linear program:

$$\min \left(\sum_{i=1}^{d} u_i \right) \quad \text{subject to } -\mathbf{u} \leq \boldsymbol{\zeta} \leq \mathbf{u}$$

$$\text{and } -\lambda_d \sigma \mathbf{1} \leq \mathbf{X}_{\mathcal{M}}^T (\mathbf{y} - \mathbf{X}_{\mathcal{M}} \boldsymbol{\zeta}) \leq \lambda_d \sigma \mathbf{1}$$

Methods of model selection with ultrahigh dimensionality

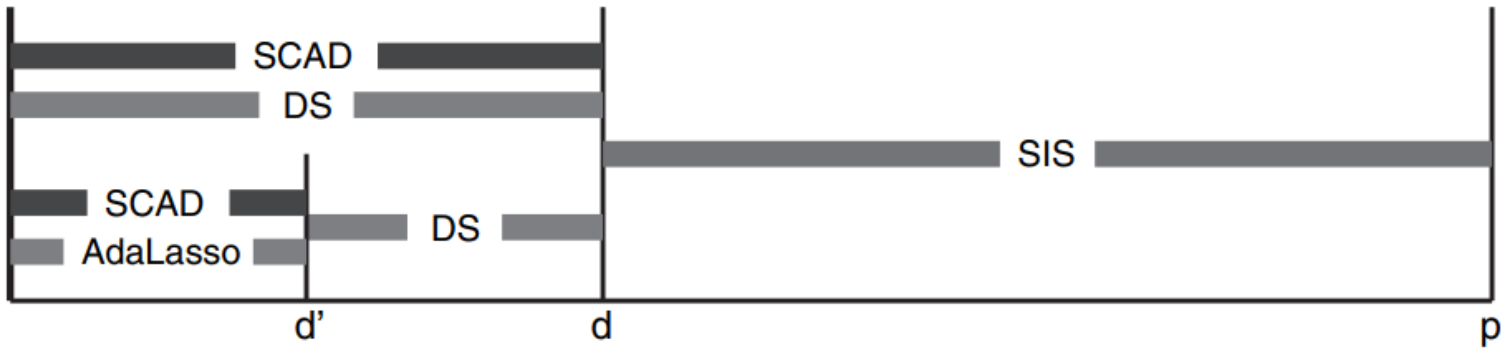


Fig. 3. Methods of model selection with ultrahigh dimensionality

Simulation 1:'independent features'

For the first simulation, we used the linear model (1) with IID standard Gaussian predictors and Gaussian noise with standard deviation $\sigma = 1.5$. We considered two such models with $(n, p) = (200, 1000)$ and $(n, p) = (800, 20000)$.The sizes s of the true models,i.e. the numbers of non-zero coefficients, were chosen to be 8 and 18, and the non-zero components of the p-vector β were randomly chosen as follows. We set $a = 4\log(n)/n^{1/2}$ and $5\log(n)/n^{1/2}$ respectively and picked non-zero coefficients of the form $(-1)^u(a + |z|)$ for each model, where u was drawn from a Bernoulli distribution with parameter 0.4 and z was drawn from the standard Gaussian distribution.

Simulation 1:'independent features'

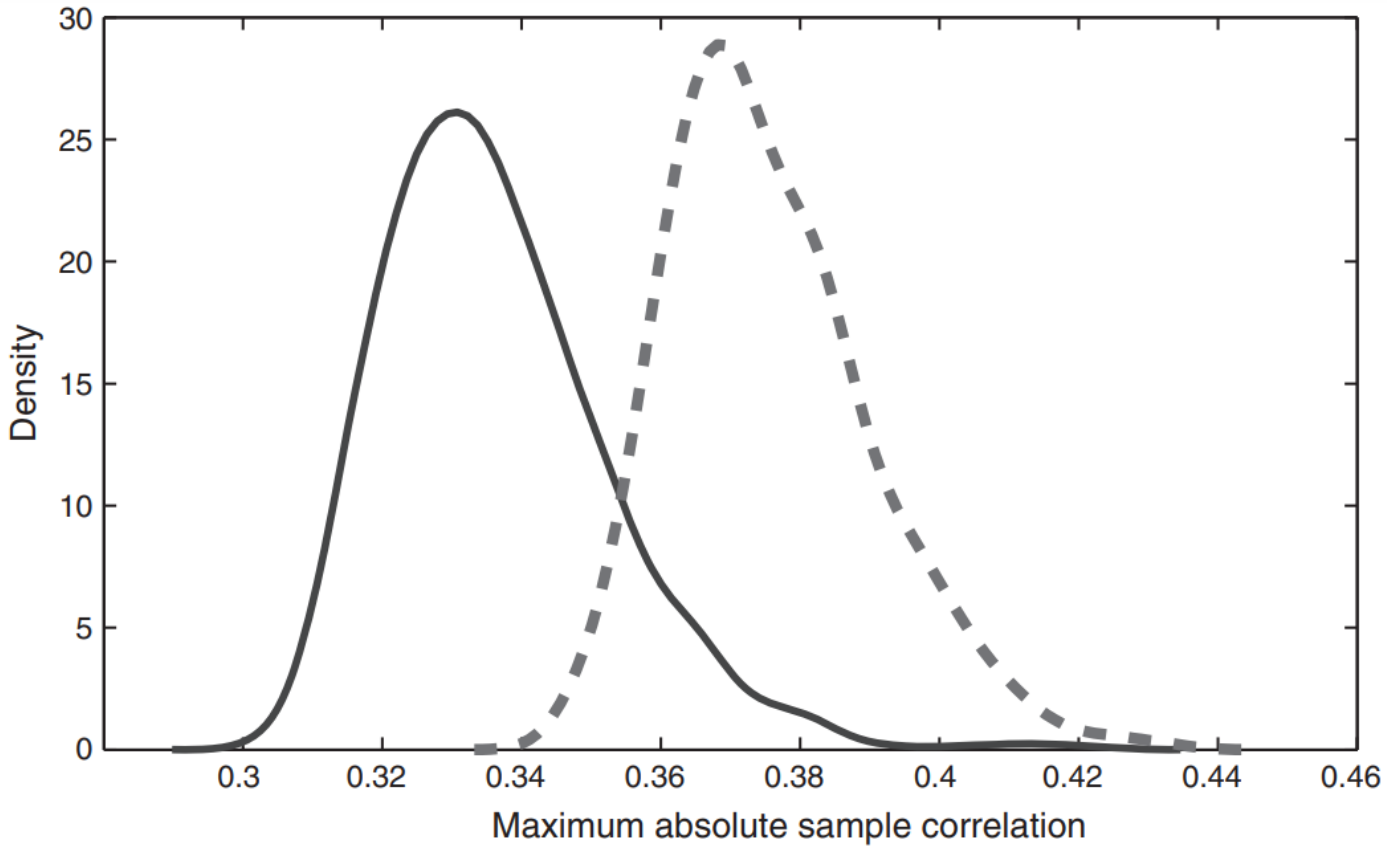


Fig. 4. Distributions of the maximum absolute sample correlation when $n = 200$ and $p = 1000$ (———) and $n = 200$ and $p = 5000$ (- - - - -)

Simulation 1:'independent features'

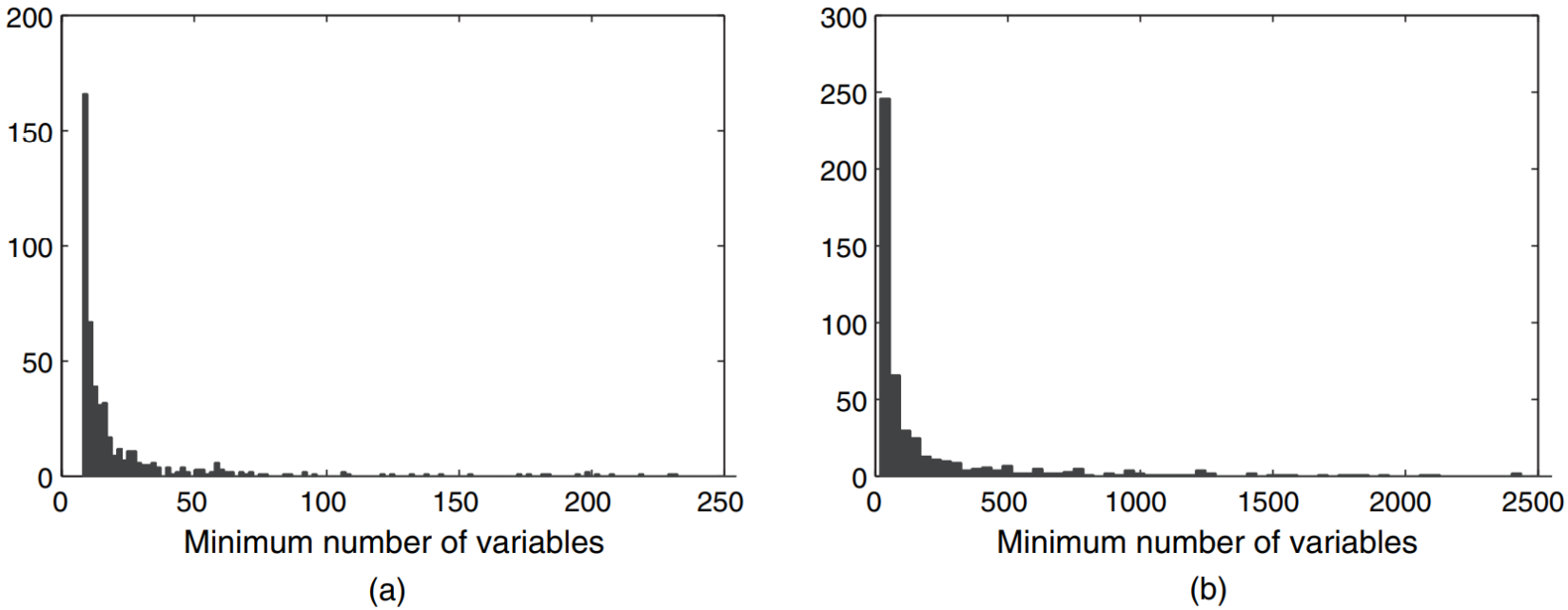


Fig. 5. Distribution of the minimum number of selected variables that is required to include the true model by using SIS when (a) $n = 200$ and $p = 1000$ and (b) $n = 800$ and $p = 20\,000$ in simulation I

Simulation 1:'independent features'

Table 1. Results of simulation I: medians of the selected model sizes and estimation errors (in parentheses)

p	*Results for the following methods:*					
	Dantzig selector	*Lasso*	*SIS–SCAD*	*SIS–DS*	*SIS–DS–SCAD*	*SIS–DS–AdaLasso*
1000	10^3	62.5	15	37	27	34
	(1.381)	(0.895)	(0.374)	(0.795)	(0.614)	(1.269)
20000	—	—	37	119	60.5	99
	—	—	(0.288)	(0.732)	(0.372)	(1.014)

Simulation 2:'dependent' features

$(n, p, s) = (200, 1000, 5), (200, 1000, 8)$ and $(800, 20000, 14)$, where s denotes the size of the true model,i.e. the number of non-zero coefficients. The three p-vectors β were generated in the same way as in simulation 1. We set $(\sigma, a) = (1, 2\log(n)/n^{1/2}), (1.5, 4\log(n)/n^{1/2}), (2, 4\log(n)/n^{1/2})$. We first used MATLAB function **sprandsym** to generate randomly an $s \times s$ symmetric positive definite matrix $\mathbf{A}$ with condition number $n^{1/2}/\log(n)$ and drew samples of s predictors $X_1, \cdot, X_s$ from $\mathcal{N}(\mathbf{0}, \mathbf{A})$.Then we took $Z_{s+1}, \cdots, Z_p \sim \mathcal{N}(\mathbf{0}, I_{p-s})$ and defined the remaining predictors as $X_i = Z_i + rX_{i-s}, i = s+1, \cdots, 2s$ and $X_i = Z_i + (1-r)X_1, i = 2s+1, \cdots, p$, with $r = 1 - 4\log(n)/p, 1 - 5\log(n)/p$ and $1 - 5\log(n)/p$. For each model we simulated 200 date sets.

Leukaemia data analysis

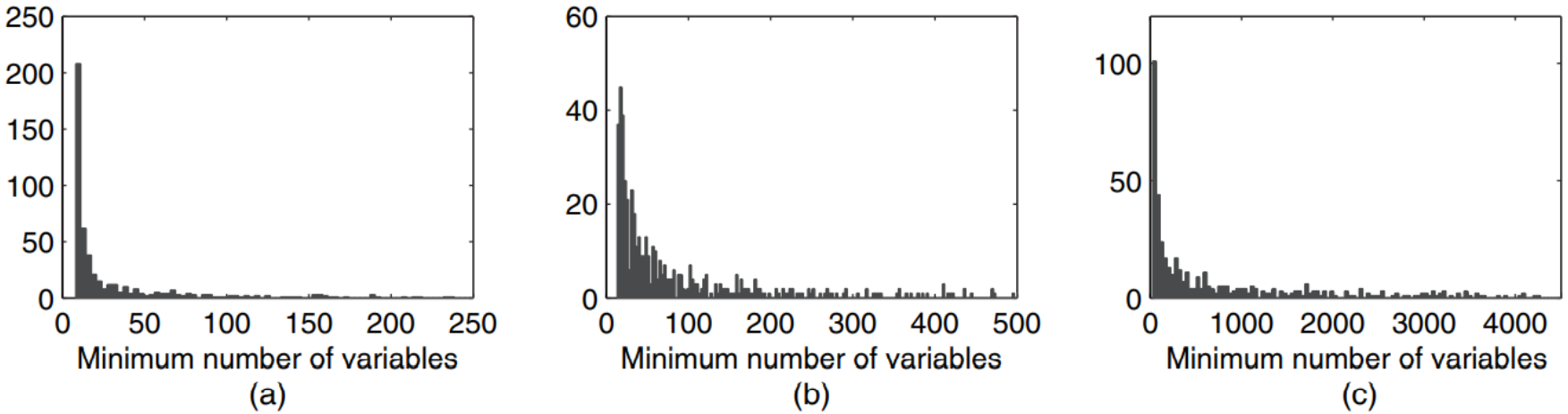


Fig. 6. Distribution of the minimum number of selected variables that is required to include the true model by using SIS when (a) $n = 200$, $p = 1000$ and $s = 5$, (b) $n = 200$, $p = 1000$ and $s = 8$ and (c) $n = 800$, $p = 20\,000$ and $s = 8$ in simulation II

Leukaemia data analysis

Table 2. Results of simulation II: medians of the selected model sizes and estimation errors (in parentheses)

p	*Results for the following methods:*					
	Dantzig selector	*Lasso*	*SIS–SCAD*	*SIS–DS*	*SIS–DS–SCAD*	*SIS–DS–AdaLasso*
1000	10^3	91	21	56	27	52
($s=5$)	(1.256)	(1.257)	(0.331)	(0.727)	(0.476)	(1.204)
($s=8$)	10^3	74	18	56	31.5	51
	(1.465)	(1.257)	(0.458)	(1.014)	(0.787)	(1.824)
20000	—	—	36	119	54	86
	—	—	(0.367)	(0.986)	(0.743)	(1.762)

Leukaemia data analysis

Table 3. Classification errors in the leukaemia data set

Method	*Training error*	*Test error*	*Number of genes*
SIS–SCAD–LD	0/38	1/34	16
SIS–SCAD–NB	4/38	1/34	16
Nearest shrunken centroids	1/38	2/34	21

Overview

Some extensions of correlation learning

The key idea of SIS is to apply a single componentwise regression. Three potential issues ,however, might arise with this approach.

1. some unimportant predictors that are highly correlated with the important predictors can have higher priority for being selected by SIS than other important perdictors that are relatively weakly related to the response.
2. 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.
3. The issue of collinearity between predictors adds difficulty to the problem of variable selection.

These three issues will be addressed in the extensions of SIS below, which allow us to use more fully the joint information of the covariates rather than just the marginal information in variable selection.

ISIS Algorithm

1. In the first step, we select a subset of k_1 variables $\mathcal{A}_1 = \{X_{i_1}, \cdots, X_{i_{k_1}}\}$ using an SIS-based model selection method such as the SIS-SCAD or SIS-lasso methods. There variables were selected, using SCAD or the lasso, on the basis of the joint information of $[n/\log(n)]$ variables that survive after correlation screening. Then we have an n-vector of residuals from regressing the response Y over $X_{i_1}, \cdots, X_{i_k}$.
2. In the next step, we treat those residuals as the new responses and apply the same method as in the previous step to the remaining $p - k_1$ variables, which results in a subset of k_2 variables $\mathcal{A}_2 = \{X_{j_1}, \cdots, X_{j_{k_2}}\}$.
3. We can keep on doing this until we obtain l disjoint subsets $\mathcal{A}_1, \cdots, \mathcal{A}_l$ whose union $\mathcal{A} = \cup_{i=1}^{l} \mathcal{A}_i$ has a size d, which is less than n. In practical implementation,we can choose , for example, the largest l such that $|\mathcal{A}| < n$. Form the selected features in $\mathcal{A}$, we can choose the features by using a moderate scale method such as SCAD, the lasso or the Dantzig selector.

Guouping and transformation of the input variables

- A notorious difficulty of variable selection lies in the collinearity between the covariates.
- A good idea is to transform the input variables.
- Subject-related transformation is a useful tool. In some cases, a simple linear transformation of the input variables.
- statistical transformation: clustering algorithm $\longrightarrow$ sparse principal components analysis $\longrightarrow$ SIS-based model.

Simulated example 1

$$Y = 5X_1 + 5X_2 + 5X_3 + \varepsilon$$

Where $X_1, \cdots, X_p$ are p predictors and $\varepsilon \sim N(0,1)$ is noise that is independent of the predictors. In the simulation, a sample of $(X_1, \cdots, X_p)$ with size n was drawn from a multivariate normal distribution $N(0, \Sigma)$ whose covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ has entries $\sigma_{ii} = 1, i = 1, \cdots, p$ and $\sigma_{ij} = \rho, i \neq j$. We considered 20 such models characterized by (p, n, ρ) with $p = 100, 1000$, $n = 20, 50, 70$ and $\rho = 0, 0.1, 0, 5, 0.9$, and for each model we simulated 200 data sets.

Simulated example 1

Table 4. Results of simulated example I: accuracy of SIS, the lasso and ISIS in including the true model $\{X_1, X_2, X_3\}$

p	n	*Method*	*Results for the following values of* ρ:			
			$\rho=0$	$\rho=0.1$	$\rho=0.5$	$\rho=0.9$
100	20	SIS	0.755	0.855	0.690	0.670
		Lasso	0.970	0.990	0.985	0.870
		ISIS	1	1	1	1
	50	SIS	1	1	1	1
		Lasso	1	1	1	1
		ISIS	1	1	1	1
1000	20	SIS	0.205	0.255	0.145	0.085
		Lasso	0.340	0.555	0.556	0.220
		ISIS	1	1	1	1
	50	SIS	0.990	0.960	0.870	0.860
		Lasso	1	1	1	1
		ISIS	1	1	1	1
	70	SIS	1	0.995	0.97	0.97
		Lasso	1	1	1	1
		ISIS	1	1	1	1

Simulated example 2

ρ was fixed to be 0.5 for simplicity

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + \varepsilon$$

where $X_4 \sim N(0,1)$ and has correlation $\rho^{1/2}$ with all the other $p-1$ variables.

Table 5. Results of simulated example II: accuracy of SIS, the lasso and ISIS in including the true model $\{X_1, X_2, X_3, X_4\}$†

p	*Method*	*Results for the following values of n:*		
		$n=20$	$n=50$	$n=70$
100	SIS	0.025	0.490	0.740
	Lasso	0.000	0.360	0.915
	ISIS	1	1	1
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	1	1	1

†$\rho = 0.5$.

Simulated example 3

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + X_5 + \varepsilon$$

Table 6. Results of simulated example III: accuracy of SIS, the lasso and ISIS in including the true model $\{X_1, X_2, X_3, X_4, X_5\}$†

p	*Method*	*Results for the following values of n:*		
		$n = 20$	$n = 50$	$n = 70$
100	SIS	0.000	0.285	0.645
	Lasso	0.000	0.310	0.890
	ISIS	1	1	1
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	1	1	1

†$\rho = 0.5$.

Simulated example 3

Table 7. Simulations I and II in Section 3.3 revisited: medians of the model sizes selected and the estimation errors (in parentheses) for the ISIS–SCAD method

p	*Results for simulation I*		*Results for simulation II*
1000	13 (0.329)	$(s=5)$	11 (0.223)
		$(s=8)$	13.5 (0.366)
20000	31 (0.246)		27 (0.315)

Overview

Some notaions

Recall from model (1) that $Y = \sum_{i=1}^{p} \beta_i X_i + \varepsilon$. We let $\mathcal{M}_* = \{1 \leq i \leq p : \beta_i \neq 0\}$ be the true sparse model with non-sparsity size $s = |\mathcal{M}_*|$ and define

$$\mathbf{z} = \mathbf{\Sigma}^{-1/2}\mathbf{x}, \mathbf{Z} = \mathbf{X}\mathbf{\Sigma}^{-1/2} \tag{13}$$

where $\mathbf{x} = (X_1, \cdots, X_p)^T$ and $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{x})$. Clearly, the n rows of the transformed design matrix $\mathbf{Z}$ are IID copies of $\mathbf{z}$ which now has covariance matrix I_p. For simplicity, all the predictors $X_1, \cdots, X_p$ are assumed to be standardized to have mean 0 and standard deviation 1. Note that the design matrix $\mathbf{X}$ can be factored into $\mathbf{Z}\mathbf{\Sigma}^{1/2}$. Below we shall make assumptions on $\mathbf{Z}$ and $\mathbf{\Sigma}$ separately.

We denote by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ the largest and smallest eigenvalues of a matrix respectively. For $\mathbf{Z}$, we are concerned with a concentration property of its extreme singular values as follows.

Concentration property

Definition of Concentration property

The random matrix $\mathbf{Z}$ is said to have the concentration property if there are some $c, c_1 > 1$ and $C_1 > 0$ such that the deviation inequality

$$P\{\lambda_{\max}(\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T) > c_1 \text{ or } \lambda_{\min}(\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T) < 1/c_1\} \leq \exp(-C_1 n) \quad (14)$$

holds for any $n \times \widetilde{p}$ submatrix $\widetilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $cn < \widetilde{p} \leq p$. We shall call it property C for short.

Property C amounts to a distributional constraint on $\mathbf{z}$. Intuitively, it means that with large probability the n non-zero singular values of the $n \times \widetilde{p}$ matrix $\widetilde{\mathbf{Z}}$ are of the same order, which is reasonable since $\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T$ will approach I_n as $\widetilde{p} \to \infty$. It relies on random-matrix theory to derive the deviation equlity (14). In particular, property C holds when $\mathbf{x}$ has a p-variate Gaussian distribution. We conjecture that it should be shared by a wide calss of spherically symmetric distributions.

Assumptions

Some of the assumptions below are purely technical and serve only to provide theoretical understanding of the newly proposed methodology. We have no intent to make our assumptions the weakest possible.

Condition 1

$p > n$ and $\log(p) = O(n^{\xi})$ for some $\xi \in (0, 1-2\kappa)$, where κ is given by condition 3.

Condition 2

$\mathbf{z}$ has a spherically symmetric distribution and property C. Also, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.

Condition 3

$\mathrm{var}(Y) = O(1)$ and, for some $\kappa \geq 0$ and $c_2, c_3 > 0$

$$\min_{i \in \mathcal{M}_*} |\beta_i| \geq \frac{c_2}{n^k} \quad \text{and} \quad \min_{i \in \mathcal{M}_*} \left|\mathrm{cov}(\beta_i^{-1} Y, X_i)\right| \geq c_3$$

Assumptions

As seen later, κ controls the rate of probability error in recovering the true sparse model. Although $b = \min_{i \in \mathcal{M}_*} \left|\mathrm{cov}(\beta_i^{-1} Y, X_i)\right|$ is assumed here to be bounded away from 0, our asymptotic study applies as well to the case where $b \to 0$ as $n \to \infty$. In particular, when the variables in $\mathcal{M}_*$ are uncorrelated, $b = 1$. This condition rules out the situation in which an important variable is marginally uncorrelated with Y, but jointly correlated with Y.

Condition 4

There are some $\tau \geq 0$ and $c_4 > 0$ such that

$$\lambda_{\max}(\mathbf{\Sigma}) \leq c_4 n^{\tau}$$

This condition rules out the case of strong collinearity.

Sure screening property

Analysing the p-vector $\boldsymbol{\omega}$ in equation (2) when $p > n$ is essentially difficult. The approach that we took is first to study the specific case with $\Sigma = I_p$ and then to relate the general case the the specific case.

Theorem 1 (accuracy of SIS)

Under conditions 1-4, if $2\kappa + \tau < 1$ then there is some $\theta < 1 - 2\kappa - \tau$ such that, when $\gamma \sim cn^{-\theta}$ with $-c > 0$, we have, for some $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O\left[\exp\left\{-Cn^{1-2\kappa}/\log(n)\right\}\right]$$

Theorem 1 shows that SIS has the sure screening property and can reduce from exponentially growing dimension p down to a relatively large scale $d = [\gamma n] = O(n^{1-\theta}) < n$ for some $\theta > 0$, where the reduced model $\mathcal{M} = \mathcal{M}_\gamma$ still contains all the variables in the true model with an overwhelming probability. In particular, we can choose the submodel size d to be $n - 1$ or $n/\log(n)$ for SIS if conditions 1-4 are satisfied.

asymptotic sure screening

The proof of theorem 1 depends on the iterative application of the following theorem, which demonstrates the accuracy of each step of ITRRS. We first describe the result of the first step of ITRRS. It shows that, as long as the ridge parameter λ is sufficiently large and the percentage of remaining variables δ is sufficiently large, the sure screening property is ensured with overwhelming probability.

Theorem 2(asymptotic sure screening)

Under conditions 1-4, if $2\kappa + \tau < 1, \lambda(p^{3/2}n)^{-1} \to \infty$, and $\delta n^{1-2\kappa-\tau} \to \infty$ as $n \to \infty$, then we have, for some $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}_{\delta,\lambda}^1) = 1 - O\left[\exp\left\{-Cn^{1-2\kappa}/\log(n)\right\}\right]$$

Theorem 2 reveals that, when the tuning parameters are chosen appropriately, with an overwhelming probability the submodel $\mathcal{M}_{\delta,\lambda}^1$ will contain the true model $\mathcal{M}_*$ and its size is an order n^θ(for some $\theta > 0$) lower than the original one. This property stimulated us to propose ITRRS.

accuracy of ITRRS

Theorem 3 (accuracy of ITRRS)

Let the assumptions of theorem 2 be satisfied. If $\delta n^{\theta} \to \infty$ as $n \to \infty$ for some $\theta < 1 - 2\kappa - \tau$, then successive applications of procedure (7) for k times results in a submodel $\mathcal{M}_{\delta,\lambda}$ with size $d = [\delta^k p] < n$ such that, for some $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}_{\delta,\lambda}) = 1 - O\left[\exp\left\{-Cn^{1-2\kappa}/\log(n)\right\}\right]$$

Theorem 3 follows from iterative application of theorem 2 k times, where k is the first integer such that $[\delta^k p] < n$. This implies that $k = O\{\log(p)/\log(n)\} = O(n^{\xi})$. Therefore, the accumulated error probability, from the union bound, is still exponentially small with a possibility of a different constant C.

ITRRS has now been shown to have the sure screening property. As mentioned before, SIS is a specific case of ITRRS with an infinite regularization parameter and hence enjoys also the sure screening property.

Consistency of methods SIS-DS

Theorem 4 (consistency of method SIS-DS)

Assume, with large probability, that $\delta_{2s}(\mathbf{X}_{\mathcal{M}}) + \theta_{s,2s}(\mathbf{X}_{\mathcal{M}}) \leq t < 1$ and choose $\lambda_d = \{2\log(d)\}^{1/2}$ in problem

$$\min_{\boldsymbol{\zeta} \in \mathbb{R}^d} (\|\boldsymbol{\zeta}\|_1) \quad \text{subject to } \|\mathbf{X}_{\mathcal{M}}^T \mathbf{r}\|_\infty \leq \lambda_d \sigma$$

Then, with large probability, we have

$$\|\hat{\beta}_{\mathrm{DS}} - \beta\|^2 \leq C\{\log(d)\}^{1/2} s\sigma^2$$

where $C = 32/(1-t)^2$ and s is the number of non-zero components of β.

Consistency of methods SIS-SCAD

Theorem 5 (oracle property of method SIS-SCAD)

If $d = o(n^{1/3})$ and the assumptions of theorem 2 in Fan and Peng (2004) are satisfied, then , with probability tending to 1, the SCAD-PLS estimator $\hat{\beta}_{\mathrm{SCAD}}$ satisfies
(a) $\hat{\beta}_i = 0$ for any $i \notin \mathcal{M}_*$ and
(b) the components of $\hat{\beta}_{\mathrm{SCAD}}$ in $\mathcal{M}_*$ perform as well as if the true model $\mathcal{M}_*$ were known.

Proof of theorem 1

- Step 1: define a submodel

$$\tilde{\mathcal{M}}_{\delta}^{1}=\{1\leq i\leq p:|\omega_{i}|\text{ is among the first }[\delta p]\text{ largest of all}\} \quad (15)$$

We aim to show that, if $\delta\to 0$ in such a way that $\delta n^{1-2\kappa-\tau}\to\infty$ as $n\to\infty$, we have, for some $C>0$,

$$P(\mathcal{M}_{*}\subset\tilde{\mathcal{M}}_{\delta}^{1})=1-O\left[\exp\left\{-Cn^{1-2\kappa}/\log\left(n\right)\right\}\right] \quad (16)$$

The main idea is to relate the general case to the specific case with $\mathbf{\Sigma}=I_{p}$

$$\boldsymbol{\omega}=\mathbf{X}^{T}\mathbf{X}\boldsymbol{\beta}+\mathbf{X}^{T}\boldsymbol{\varepsilon}:=\boldsymbol{\xi}+\boldsymbol{\eta}$$

Proof of theorem 1

- Step 1.1: first, we consider term $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_p)^T = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$
- Step 1.1.1 (bounding $\|\boldsymbol{\xi}\|$ form above)
- Step 1.1.2 (bounding $|\xi_i|, i \in \mathcal{M}_*$,from below)
- Step 1.2: then, we examine term $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_p)^T = \mathbf{X}^T\boldsymbol{\varepsilon}$
- Step 1.2.1 (bounding $\|\boldsymbol{\eta}\|$ from above)
- Step 1.2.2 (bounding $|\eta_i|$ from above)
- Step 1.3:finally, we combine the results that were obtained in steps 1.1 and 1.2.
- Step 2 fix an arbitrary $r \in (0, 1)$ and choose a shrinking factor δ of the form $(n/p)^{1/(\kappa - r)}$, for some integer $k \geq 1$. We successively perform dimensionality reduction

Proof of theorem 2

We observe that expression (7) uses only the order of componentwise magnitudes of $\boldsymbol{\omega}^{\lambda}$, so it is invariant under scaling. Now we rewrite the p-vector $\lambda\boldsymbol{\omega}^{\lambda}$ as

$$\lambda\boldsymbol{\omega}^{\lambda} = \boldsymbol{\omega} - \{I_p - (I_p + \lambda^{-1}\mathbf{X}^{\mathbf{X}})^{-1}\}\boldsymbol{\omega}$$

Let $\zeta = \{I_p - (I_p + \lambda^{-1}\mathbf{X}^{\mathbf{X}})^{-1}\}\boldsymbol{\omega}$ By conditions 2 and 4, and Bonferroni's inequality show that

$$P\{\|\zeta\| > O(\lambda^{-1}n^{1+3\tau}/2p^{3/2})\} \leq O[s\exp\{-Cn^{1-2\kappa/\log(n)}\}]$$

Note that $\kappa + \tau/2 < \frac{1}{2}$ by assumption. So in particular, we can choose any λ satisfying $\lambda(p^{3/2}n)^{-1} \to \infty$ as $n \to \infty$.

Proof of theorem 3

Theorem 3 is a straightforward corollory to theorem 2 by the argument in step 2 of the proof of theorem 1. The distribution of $\mathbf{z}$ is continuous and spherically symmetric,i.e. invariant under the orthogonal group $\mathcal{O}(p)$.

Overview

Introduction

- The key idea of the SIS procedure is to rank all predictors by using a utility measure between the response and each predictor and then to retain the top variables for further investigation. The SIS procedure has been rapidly extended to various models and data types.

Introduction

- Zhu et al.(2011) used the expectation of the square of the correlation between the predictor and an indicator function of the response for an ultra-high-dimensional multi-index model (SIRS), and Li et al.(2012) used distance correlation to carry out marginal screening (DC-SIS)
- Shao and Zhang (2014) proposed a martingale difference correlation for high-dimensional variable screening (MDC-SIS). For variable interaction, Fan et al. (2014) proposed a sure independent screening procedure based on Pearson correlation (P-IT), and Kong et al. (2014) developed one based on distance correlation (DC-IT).

Introduction

- With the availability of more data types and possible models, a model-free generic screening procedure with fewer and less restrictive assumptions is desirable. We propose a generic nonparametric sure independence screening procedure, called BCor-SIS, on the basis of a recently developed universal dependence measure: Ball correlation.
- We investigate the flexibility of this procedure by considering three commonly encountered challenging settings in biological discovery or precision medicine: iterative BCor-SIS, interaction pursuit, and survival outcomes.

Definition

The Ball covariance $\mathbf{BCov}(X, Y)$ is defined as the square root of

$$\mathbf{BCov}^2(X, Y) = \int\int_{U\times V} [\theta - \mu \otimes \nu]^2 (\overline{B}_{\zeta X}(x_1, x_2) \times \overline{B}_{\zeta Y}(y_1, y_2)) \\ \theta(dx_1, dy_1)\theta(dx_2, dy_2)$$

where $\mu \otimes \nu$ is a product measure on $\mathcal{X} \times \mathcal{Y}$.

Methodologies

Definition

The Ball correlation $\mathbf{BCov}(X, Y)$ is defined as the square root of

$$\mathbf{BCor}^2(X, Y) = \mathbf{BCov}^2(X, Y)/\sqrt{\mathbf{BCov}^2(X, X) \times \mathbf{BCov}^2(Y, Y)}$$

if $\mathbf{BCov}(X, X) \times \mathbf{BCov}(Y, Y) > 0$, or 0 otherwise.

Let $\delta^X_{ij,k} := I(X_k \in \overline{B}_{\zeta_X}(X_i, X_j))$

Definition

Empirical Ball covariance $\mathbf{BCov}_n(\mathbf{X}, \mathbf{Y})$ is defined as the square root of

$$\mathbf{BCov}^2_n(\mathbf{X}, \mathbf{Y}) = n^{-6} \sum_{i,j,k,l,s,t=1}^{n} \xi^X_{ij,klst} \xi^Y_{ij,klst}$$

Methodologies

Definition

Empirical Ball correlation $\mathbf{BCov}_n(\mathbf{X}, \mathbf{Y})$ is defined as the square root of

$$\mathbf{BCor}_n^2(X, Y) = \mathbf{BCov}_n^2(X, Y)/\sqrt{\mathbf{BCov}_n^2(X, X) \times \mathbf{BCov}_n^2(Y, Y)}$$

BCor-SIS is based on the assumption that the predictors with larger BCor are more strongly correlated with the response vector.Specifically, BCor-SIS consists of two steps:

1. Calculate $\hat{\rho}_r = \mathbf{BCor}_n^2(\mathbf{X}_r, \mathbf{Y})$, which is an estimate of $\rho_r = \mathbf{BCor}^2(\mathbf{X}_r, \mathbf{Y})$,and use it as a marginal utility of X_r for $r = 1, \cdots, \rho$.
2. Select the X_rs that fall into $\hat{\mathcal{A}}_n^* = \{r : \hat{\rho}_r \geq \tau_n, r = 1, \cdots, p\}$, where τ_n is a pre-specified constant.

Strong screening consistency of BCor-SIS

Theorem (Strong screening consistency of BCor-SIS)

There exists a positive constant $c_1 > 0$ such that

$$\mathbb{P}(\max_{1\le r\le p} |\hat{\rho}_r - \rho_r| \ge cn^{-\kappa}) \le O(p \times \exp(-c_1 n^{1-2\kappa}))$$

If condition $(\mathrm{C1})$ holds and (X_r, Y) satisfies the conditions of Lemma 1, then for any $\tau_n \in (0, 2cn^{-\kappa})$, there exists a constant $c_2 > 0$ such that

$$\mathbb{P}(\mathcal{A} \subset \hat{\mathcal{A}}_n^*) \ge 1 - O(\gamma \exp(-c_2 n^{1-2\kappa}))$$

$$\mathbb{P}(\hat{\mathcal{A}}_n^* \subset \mathcal{A}) \ge 1 - O(\gamma^* \exp(-c_2 n^{1-2\kappa}))$$

References

- Fan, Jianqing, and Jinchi Lv. "Sure independence screening for ultrahigh dimensional feature space." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70.5 (2008): 849-911.
- Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American statistical Association 96.456 (2001): 1348-1360.
- Candes, Emmanuel, and Terence Tao. "The Dantzig selector: Statistical estimation when p is much larger than n." The annals of Statistics 35.6 (2007): 2313-2351.
- Zou, Hui. "The adaptive lasso and its oracle properties." Journal of the American statistical association 101.476 (2006): 1418-1429.
- Pan, Wenliang, et al. "A generic sure independence screening procedure." Journal of the American Statistical Association (2018).