Sure Independence Screening for Ultra-High Dimensional Feature Space

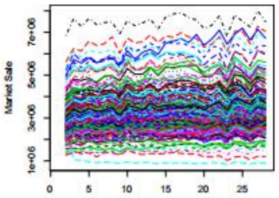
presenter: Fei Yang 2021/04/30

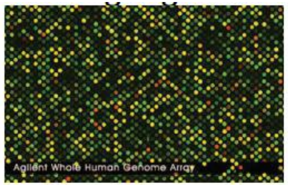
CONTENTS

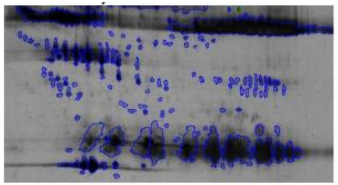
1.1Background information

Variable selction plays an important role in high dimensional statistical modeling. Frequent in:

- **Biological science:** disease classification/ predicting clinical outcomes using high-throught data; association studies;
- **Engineering:** Doc or text classification, computer vision;
- **Economics, Finance, Marketing:** sale data collected in many regions
- **Spatial-temporal:Meteorology:** Earth Sciences; Ecology.







1.1Background information

- While adding much greater flexibility to modeling with enriched feature space, ultrahigh-dimensional data analysis poses fundamental challenges to scalable learning and inference with good statistical efficiency.

- Sure independence screening(SIS)确立性独立筛选 is a simple and effective method to this endeavor. This framework of two-scale statistical learning introduced in Fan and Lv (2008), has been extended to various model settings ranging

- Parametric
- Semiparametric
- Nonparametric

- regression
- classification
- survival analysis

1.2 Insight into high dimensionality

Consider the variable selection problem in linear model

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T, \beta = (\beta_1, \cdots, \beta_p)^T, \varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T, and\ X = (x_1, \cdots, x_p)^T, \sum = cov(x), z = \sum^{-\frac{1}{2}} x$

$\varepsilon_i\ \ i.i.d\ mean\ 0\ and\ variance\ \sigma^2$

$Y_i\ centered$

$X_i\ standardized$

当p>>n时,估计β过程中会遇到的问题有:

1. $X^T X$非列满秩，OLS估计非一致性

2.X存在多重共线性时，$X^T X$趋近于0，OLS估计不存在

3.最小非零的$|\beta_i|$可能会随着n的增大而衰减到噪声水平

$X_{ij}\beta_j$相对于模型误差ε_i很小且模型的信噪比较大(SNR=$\frac{var(X_i\beta)}{var(\varepsilon_i)}$)

4.z可能为厚尾分布

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \Longrightarrow \hat{\beta}^{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$$

1.2 Insight into high dimensionality

Consider the variable selection problem in linear model

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T, \beta = (\beta_1, \cdots, \beta_p)^T, \varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T, and\ X = (x_1, \cdots, x_p)^T, \sum\ = cov(x), z = \sum^{-\frac{1}{2}} x$

$\varepsilon_i\ \ i.i.d\ mean\ 0\ and\ variance\ \sigma^2$

$Y_i\ centered$

$X_i\ standardized$

一种解决方案：

以岭回归方法为代表的惩罚函数回归方法。

岭回归法的基本原理是：

在限定系数向量的l2范数大小情况下，使残差平方和最小。当自变量之间存在多重相关性时，岭回归提供一个比OLS更稳定的估计，回归系数标准差更小，通过使bias和variance的组合效应达到最佳水平，同时提高回归模型的稳定性和预测精度

1.3 Existing Methods

基于惩罚函数的变量筛选方法。

Method	Evaluation	YEAR	Target Function
AIC,BIC, best subset selection	Combinatoric, NP-hard problem, computational intensive when p is large		
LASSO	Provide sparsity solution, model selection consistency: very strong conditions(Zhang and Yu (2006))	1996	$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2 + \lambda\sum_{i=1}^{d}\lvert\beta_j\rvert$
Bridge	Provide sparsity solution(0<q<1) , includes LASSO(l_1) and Ridge(l_2) as special case(l_q)(q>0)	1993	$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2 + \lambda\sum_{i=1}^{d}\lvert\beta_j\rvert^{\gamma}$ where $0<\gamma<1$
SCAD	Oracle property, low dimension $\frac{p^3}{n} \to 0$(Fan and Peng 2004)	2001	$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2 + \sum_{i=1}^{p} p_j(\lvert\beta_j\rvert)$
Adaptive LASSO	Oracle property, low dimension (Zou 2006)	2006	$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2 + \lambda_n\sum_{i=1}^{p} w_j\lvert\beta_j\rvert$
Dantzig Selector	High dimension (p>n), Oracle property Need uniform uncertainty principle condition (UUP)(Candes and Tao 2007). Linear programming is slow in ultrahigh dimension. p can not grow exponentially	2007	$min\ \lvert\lvert X^T(Y_i - X_i'\beta)\rvert\rvert_\infty + \lambda\sum_{i=1}^{d}\lvert\beta_j\rvert$

1.3 Existing Methods

基于惩罚函数的变量筛选方法。

Method	Evaluation	YEAR	Target Function
AIC,BIC, best subset selection	Combinatoric, NP-hard problem, computational intensive when p is large		
LASSO	优点：计算复杂度小，且参数估计具有连续性 缺点：相合性不好	1996	$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^{'}\beta)^2 + \lambda\sum_{i=1}^{d}\|\beta_j\|$
Bridge	Provide sparsity solution(0<q<1) , includes LASSO(l_1) and Ridge(l_2) as special case(l_q)(q>0)	1993	$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^{'}\beta)^2 + \lambda\sum_{i=1}^{d}\|\beta_j\|^{\gamma}$ where 0<$\gamma < 1$
SCAD	Oracle property, low dimension $\frac{p^3}{n} \to 0$(Fan and Peng 2004)	2001	$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^{'}\beta)^2 + \sum_{i=1}^{p} p_j(\|\beta_j\|)$
Adaptive LASSO	Oracle property, low dimension (Zou 2006)	2006	$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^{'}\beta)^2 + \lambda_n\sum_{i=1}^{p} w_j\|\beta_j\|$
Dantzig Selector	优点：使用相关残差而非残差，有助于选取与Y高度相关的X	2007	$min\ \|\|X^T(Y_i - X_i^{'}\beta)\|\|_{\infty} + \lambda\sum_{i=1}^{d}\|\beta_j\|$

1.3 Existing Methods

LASSO类惩罚：

Relaxed Lasso（控制系数压缩速度）

Adaptive Lasso（调整不同估计量的惩罚力度）

LASSO类惩罚的拓展：

Elastic Net和Group Lasso（群组变量）、Fused Lasso（有序变量）、

Graph Lasso（图结构）。

非凸惩罚：

SCAD、MCP（大系数的近似无偏性、稀疏性、连续性）

其他类惩罚函数：

Dantzing selector（DS）及其衍生方法、SIS 及其衍生方法

Source from Qing Zhao

1.3 Existing Methods SCAD

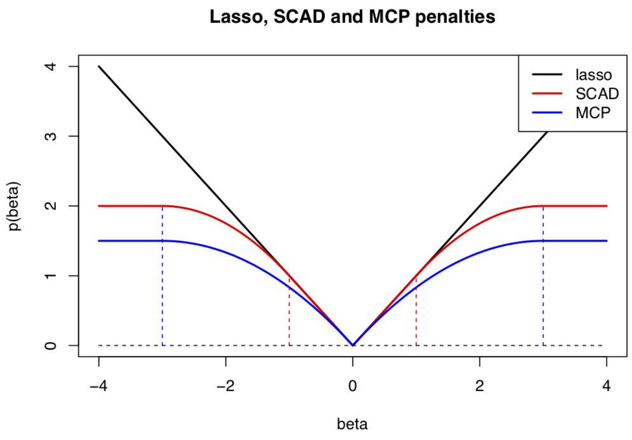
考虑到LASSO和弹性网是有偏估计，Fan and Li(2001)提出了一种连续可微的非凸惩罚函数SCAD，将很小的系数压缩到0，同时保证大系数的近似无偏性，从而降低了预测误差提高了模型精度，SCAD的惩罚力度为：

$$P(x|\lambda,\gamma)=\begin{cases}\lambda|x|, & |x|\le\lambda\\ \dfrac{2\gamma\lambda|x|-x^2-\lambda^2}{2(\gamma-1)}, & \lambda<|x|<\gamma\lambda\\ \dfrac{\lambda^2(\gamma+1)}{2}, & |x|\ge\gamma\lambda\end{cases}$$

其中，$\gamma>2$。当$|x|\le\lambda$时，该惩罚函数与 Lasso 惩罚函数一致;当$\lambda<|x|\le\gamma\lambda$时,用一个凹的二次函数进行惩罚(随着$|x|$的增大,惩罚力度逐渐减少);当$|x|\ge\gamma\lambda$时,用一个常数进行惩罚;

- SCAD对系数估计量的惩罚速度随着系数估计量绝对值的增大而逐渐减少
- SCAD在原点的导数存在，保证了稀疏性和连续性．
- 在高维数据下，SCAD估计量在一定条件下，具有oracle property。

Source from Qing Zhao



1.3 Existing Methods

Dantzig Selector(DS)

DS方法的参数估计为下述凸优化问题的解：

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

$Y_i - X_i^{'}\xi$

$\lambda_d\sigma$ 是协调参数，σ是真模型误差的标准差

$\lambda_d\sigma$*建议采用固定的调整参数*

$\lambda_d\sigma=(1+t^{-1})\sigma\sqrt{2logp}$ (t>0)

优点:

1. 在低维度方法中，且满足UUP（Uniform Uncertainty Principle）一致不确定原则下，DS方法对参数估计的误差有很好的控制

缺点:

1. 在高维场景下计算复杂度较高
2. 协调参数中的log p会受到维数影响
3. 高维下，无法满足UUP条件，模型误差无法满足正态性假设
4.无法保证选出right model

2.1 Sure Independence Screening

Sure independence screening: By using **correlation ranking** $r_i = |\mathrm{corr}(X_i, Y)|$ (Fan and Lv, 2008),

★ reduce dim from $p = O(\exp(n^a))$ to $d = o(n)$

★ Limitations: ■Linear models. ■Joint normality.

$$Y = \sum_{j \in \mathcal{M}_*} \beta_j X_j + \varepsilon$$

- Fan and Lv(2008) 首次提出了超高维变量筛选（Variable Selection）的概念，对线性模型协变量和响应变量的Pearson相关系数进行了详细的理论分析，建立了确保筛选性质，提出了确保独立筛选（Sure Independence Screening，SIS）方法和迭代确保独立筛选方法（Iterative Sure Independence Screening，ISIS），将超高维数p压缩到适当的维数d(d≤n)。

- 超高维变量选择的主要方法是通过Pearson 相关分析建立确保筛选性质，选择重要变量，然后由惩罚方法实现变量选择和参数估计。

2.2 SIS-Overview

Correlation Learning

Uses marginal correlation of features to the response variable to rank their importance

Low Computational Cost

O(np)

Sure Screening

Probability that all important variables survive is 1

Given $\mathcal{M}_*$ the true model and $\mathcal{M}_\gamma$ the model selected by SIS:

Theorem

$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1$ as $n \to \infty$

Sure Independence Screening

- A broader variation of correlation learning
  - Ranks the importance of features according to the marginal correlation with the response variables

•Reduce logrithmic factor:

Log(p) ⟶Log(d)< Log(n)

•Oracle Property神谕性:

Selecting right model; estimating parameters efficiently

2.3 SIS-Methodology

- Consider a linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

Where $\beta = (\beta_1, \beta_2 \ldots\ldots, \beta_p)^T$ is p ×1 vector of parameters, (p>n)

- When p>>n， the least square of β ($\hat{\beta} = (X^T X)^{-1} X^T Y$) is not well defined due to the singularity of $X^T X$
- A useful technique to deal with singularity of the design matrix X is the ridge regression , defined by

$$\widehat{\beta_\lambda} = (X^T X + \lambda Ip)^{-1} X^T Y$$

Where λ is a ridge parameter

- If $\lambda \to 0$, then $\widehat{\beta_\lambda}$ tends to be the least squares estimator
- If $\lambda \to \infty$， then $\lambda\widehat{\beta_\lambda}$ tends to $X^T Y$
- This implies that $\widehat{\beta_\lambda} \propto X^T Y$

2.3 SIS-Methodology-Pearson correlation

Pearson边际相关

- Consider a linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

- In practice, all covariates and the response are marginally standardized respectively ($\mu = 0,\ \sigma^2 = 1$)
- Then $\frac{1}{n}X^TY$ becomes the vector consists of the sample version of **Pearson correlations** between the response and individual covariate. This is the motivation of using Pearson correlation as a marginal utility for feature screening.
- Specifically denote,

$$\omega_j = \frac{1}{n}\boldsymbol{X}_j{}^T\boldsymbol{Y}, \quad for\ j = 1,2,\ldots\ldots,p$$

- Here, it is assumed that both $\boldsymbol{X}_j$ and $\boldsymbol{Y}$ are marginally standardized
- ω_j is indeed the sample correlation between the j-th predictor and the response variable

对于线性模型，将协变量与响应变量标准化，

即EY=0,EX_j=0,varY=1,var(X_j)=1,j=1,...,p 因此，

$$X^TY = X^TX\beta + X^T\varepsilon,$$

假定$E(\varepsilon_j)$=0, X_i与 $X_j(i \neq j)$正交，可得

$$\hat{\beta} = X^TY \cdot cov(X_{ij}, Y) = E(X_{ij}Y) = \omega(X_{ij}, Y)$$

其中$\omega(X_{ij}, Y)$表示随机变量X_{ij}与Y的边际相关系数，

简记为ω_j 表示第j个协变量与响应变量Y的边际相关系数。

$$\omega_j = \frac{1}{n}\boldsymbol{X}_j{}^T\boldsymbol{Y}, \quad for\ j = 1,2,\ldots\ldots,p$$

2.3 SIS-Methodology

$$Y = \begin{matrix} Y_1 \\ \vdots \\ Y_n \end{matrix} \; and \; X = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}$$

- centered and standerlized
- Then create models according to $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$
  - $\boldsymbol{Y} \sim \boldsymbol{X_1}, \cdots, \boldsymbol{X_p}$
- Then find $\widehat{\omega_j}$ for each model, i.e.$\widehat{\omega_1}, \cdots \widehat{\omega_p}$
- Then rank the absolute value of correlation $\widehat{\omega_1}, \cdots \widehat{\omega_p}$ to obtain d largest ones that d< n
- Now we have reduced the variables from p to d, we can choose from many different low-dimensional methods to reduce the parameter space further

2.3 SIS-Methodology-Pearson correlation

- Fan and Lv suggested ranking all predictors according to $|\omega_j|$
- To be specific, for any given $\gamma \in (0,1)$, the$[\gamma n]$ top ranked predictors are selected to obtain the submodel

$$\widehat{M_\gamma} = \{1 \leq j \leq p : |\omega_j| \; is\ among\ the\ first\ [\gamma n]\ largest\ of\ all\}$$

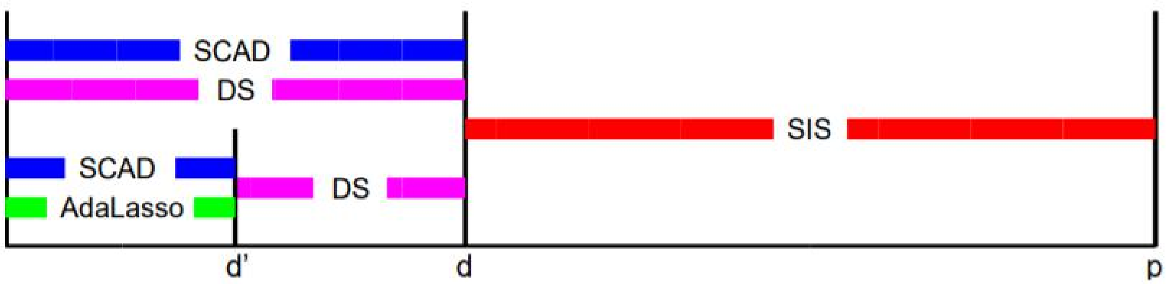


Figure 2: Methods of model selection with ultra high dimensionality.

1. Apply SIS to reduce dimensiionality from p to large scale d (d<n)
2. Use lower dimensional model selection method (SCAD, DS, AdaLasso)

2.4 SIS-Simulation 1:” independent” features

(n, p, s) = (200, 1000, 8) and (800, 20000, 18)
200 datasets

Data:
$(-1)^{u}(a+|z|)$
$u \sim Ber(0.4), z \sim N(0,1)$

n	p	s	a
200	1000	8	$4\frac{\log n}{\sqrt{n}}$
800	20000	18	$5\frac{\log n}{\sqrt{n}}$

Table 1: Results of simulation I

	Medians of the selected model sizes (upper entry) and the estimation errors (lower entry)					
p	DS	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

Computation
Limits for DS & Lasso

$d = [\frac{n}{\log n}]$

$d = n\text{-}1$
$d' = [\frac{n}{\log n}]$

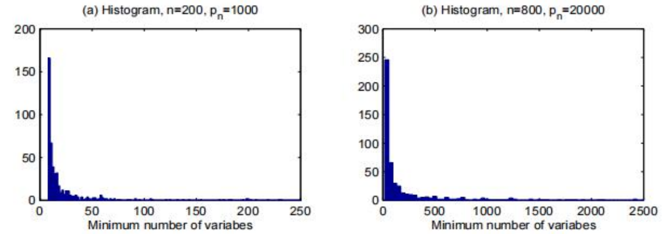


Figure 4: (a) Distribution of the minimum number of selected variables required to include the true model by using SIS when $n = 200, p = 1000$ in simulation I. (b) The same plot when $n = 800, p = 20000$.

Figure 4:SIS 保证真模型被选入的最少需要的model size

2.4 SIS-Simulation 1:” dependent” features

Data: $(-\mathbf{1})^{\boldsymbol{u}}(\boldsymbol{a}+|\boldsymbol{z}|)$

$u \sim Ber(0.4), z \sim N(0,1)$

生成给定条件数为$\frac{\sqrt{n}}{\log n}$的$s \times s$的对称正定矩阵A

取$X_1, \cdots X_s \sim N(0, A)$, 取$Z_{s+1}, \cdots Z_p \sim N(0, I_{p-s})$

定义$X_i = Z_i + rX_{i-s}$, $i = s+1, \cdots, 2s$

$X_i = Z_i + (1-r)X_1$, $i = 2s+1, \cdots, p$

(n, p, s) = (200, 1000, 5) ,(200, 1000, 8) ,(800, 20000, 14)

Table 2: Results of simulation II

p	DS	Lasso	SIS-SCAD	SIS-DS	SIS-DS-SCAD	SIS-DS-AdaLasso
	Medians of the selected model sizes (upper entry) and the estimation errors (lower entry)					
1000	10^3	91	21	56	27	52
$(s=5)$	1.256	1.257	0.331	0.727	0.476	1.204
	10^3	74	18	56	31.5	51
$(s=8)$	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

n	p	s	a	σ	r
200	1000	5	$2\frac{\log n}{\sqrt{n}}$	1	1-4 $\frac{\log n}{\sqrt{n}}$
200	1000	8	$4\frac{\log n}{\sqrt{n}}$	1.5	1-5 $\frac{\log n}{\sqrt{n}}$
800	20000	14	$4\frac{\log n}{\sqrt{n}}$	2	1-5 $\frac{\log n}{\sqrt{n}}$

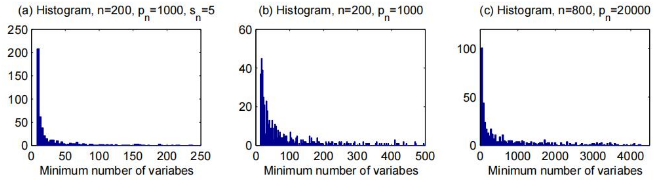


Figure 5: (a) Distribution of the minimum number of selected variables required to include the true model by using SIS when $n = 200, p = 1000, s = 5$ in simulation II. (b) The same plot when $n = 200, p = 1000, s = 8$. (c) The same plot when $n = 800, p = 20000$.

Figure 5:SIS 保证真模型被选入的最少需要的model size

2.4 SIS-Potential Issues

Potential Issues with SIS
1. Some unimportant predictors highly correlated with important predictors.(False Positive假阳性)
2.Important predictors that is marginally uncorrelated but jointly correlated with response cannot be picked.(False Negative漏诊)
3.Collinearity of predictors多重共线性

- **False Positive**: What if $X_2, \cdots, X_{99}$ highly correlated with an important X_1, but weakly correlated with Y conditionally?

$$Y = X_1 + \mathbf{0.2} X_{100} + \varepsilon$$

- **False Negative**: What if X_4 marginally uncorrelated with Y, but jointly correlated with Y?

$$Y = X_1 + X_2 + X_3 + \boldsymbol{\beta_4 X_4} + \varepsilon \quad \text{s.t.} \quad \operatorname{cov}(Y, X_4) = 0.$$

ISIS
(iterative sure independence screening)

3.1 iterive SIS(ISIS)

- Select subset of k_1 variables $A_1 = \left\{X_{i_1}, X_{i_2}, ..., X_{i_{k_1}}\right\}$
- Use n-vector of residuals as new responses and reapply SIS to remaining $p - k_1$ variables $A_2 = \left\{X_{j_1}, X_{j_2}, \ldots, X_{j_{k_2}}\right\}$
- Weaken priority of unimportant variables
- Variables missed in first screening will survive
- Stop until we get ℓ disjoint subsets of A_1, ..., A_ℓ whose union $A = \cup_{i=1}^{\ell} A_i$ has a size d, which is less than n

3.1 iterive SIS(ISIS)--Numerical Studies

For the first simulated example, we used a linear model

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

where $X_1, \cdots, X_p$ are p predictors and $\varepsilon \sim N(0,1)$ is a 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 18 such models characterized by (p, n, ρ) with $p = 100, 1000$, $n = 20, 50, 70$, and $\rho = 0, 0.1, 0.5, 0.9$, respectively, and for each model we simulated 200 data sets.

ISIS picks all true variables.

4: Results of simulated example I: Accuracy of SIS, LASSO and ISIS in including the true model $\{X_1, X_2, X_3\}$

p	n		$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
100	20	SIS	.755	.855	.690	.670
		LASSO	.970	.990	.985	.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	.205	.255	.145	.085
		LASSO	.340	.555	.556	.220
		ISIS	1	1	1	1
	50	SIS	.990	.960	.870	.860
		LASSO	1	1	1	1
		ISIS	1	1	1	1
	70	SIS	1	.995	.97	.97
		LASSO	1	1	1	1
		ISIS	1	1	1	1

3.1 iterive SIS(ISIS)--Numerical Studies

For the second simulated example, we used $\boxed{\text{the same setup}}$ as in example I except that ρ was fixed to be 0.5 for simplicity. In addition, we added a fourth variable X_4 to the model and the linear model is now

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\boxed{\sqrt{\rho}X_4} + \varepsilon,$$

where $X_4 \sim N(0, 1)$ and has correlation $\sqrt{\rho}$ with all the other $p - 1$ variables. The way X_4 was introduced is to make it uncorrelated with the response Y. Therefore, the SIS can not pick up the true model except by chance.

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

p	$\rho = 0.5$	$n = 20$	$n = 50$	$n = 70$
	SIS	.025	.490	.740
100	LASSO	.000	.360	.915
	ISIS	1	1	1
	SIS	.000	.000	.000
1000	LASSO	.000	.000	.000
	ISIS	1	1	1

ISIS alSIS picks all true variables.

3.1 iterive SIS(ISIS)--Numerical Studies

For the third simulated example, we used the same setup as in example II except that we added a fifth variable X_5 to the model and the linear model is now

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + X_5 + \varepsilon,$$

where $X_5 \sim N(0,1)$ and is uncorrelated with all the other $p-1$ variables. Again X_4 is uncorrelated

X_4与Y无关;
X_5与Y弱相关(近似误差项);

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

ISIS： Mearsures Accuracy

p	$\rho = 0.5$	$n = 20$	$n = 50$	$n = 70$
100	SIS	.000	.285	.645
	LASSO	.000	.310	.890
	ISIS	1	1	1
1000	SIS	.000	.000	.000
	LASSO	.000	.000	.000
	ISIS	1	1	1

3.2 SIS in GLM

SURE INDEPENDENCE SCREENING IN GENERALIZED LINEAR MODELS WITH NP-DIMENSIONALITY

- 目标：在广义线性模型下进行变量筛选

$$f_Y(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}$$

- 方法：通过两种方法扫描：
  - 1.By MMLE /Maximum Marginal Likelihood Estimator
  - 2. By MML/Maximum Marginal Likelihood
- 数值模拟研究
  - 逻辑回归
  - 线性模型
- 贡献：
  - 对于潜在的总体变量筛选方法，需满足1.保持真模型的非稀疏性结构；2.计算可行有效
  - 对于潜在的样本变量筛选方法，指出R方统计量、边际伪似然等也有望成为扫描依据。

3. Independence screening with MMLE. Let $\mathcal{M}_\star = \{1 \le j \le p_n : \beta_j^\star \ne 0\}$ be the true sparse model with nonsparsity size $s_n = |\mathcal{M}_\star|$, where $\boldsymbol{\beta}^\star = (\beta_0^\star, \beta_1^\star, \ldots, \beta_{p_n}^\star)$ denotes the true value. In this paper, we refer to marginal models as fitting models with componentwise covariates. The maximum marginal likelihood estimator (MMLE) $\hat{\boldsymbol{\beta}}_j^M$, for $j = 1, \ldots, p_n$, is defined as the minimizer of the componentwise regression

$$\hat{\boldsymbol{\beta}}_j^M = (\hat{\beta}_{j,0}^M, \hat{\beta}_j^M) = \arg\min_{\beta_0, \beta_j} \mathbb{P}_n l(\beta_0 + \beta_j X_j, Y),$$

where $l(Y;\theta) = -[\theta Y - b(\theta) - \log c(Y)]$ and $\mathbb{P}_n f(X,Y) = n^{-1}\sum_{i=1}^n f(X_i, Y_i)$

We select a set of variables

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \le j \le p_n : |\hat{\beta}_j^M| \ge \gamma_n\}, \tag{3}$$

where γ_n is a predefined threshold value. Such an independence learning

3.5 Rank correlation （RRCS)

ROBUST RANK CORRELATION BASED SCREENING

- 方法：利用Kentall τ 相关系数进行扫描
- 数值模拟研究
  - 线性模型
  - 广义Box-Cox变换模型
  - 逻辑回归
- 特点：
  - 可以用于半参数模型
  - Sure Independence Screening Property仅在响应变量二阶矩存在的时候具有
  - 可以用于剔除离群点与强影响点
  - indicator functions极大化简理论推导过程

2. Robust rank correlation screening (RRCS).

Consider a more general model as

$$H(Y_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.7}$$

where $\varepsilon_i, i = 1, \ldots, n$, are i.i.d. random errors independent of $\mathbf{X}_i$ with mean zero and an unknown distribution F, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-vector of parameters, its norm constrained to 1 ($\|\boldsymbol{\beta}\| = 1$) for identifiability. $H(\cdot)$ is an unspecified strictly increasing function.

For model (2.7), the invariance against any strictly increasing transformation yields that

$$\begin{aligned} \omega_k &= \frac{1}{n(n-1)} \sum_{i \neq j}^{n} I(X_{ik} < X_{jk}) I(Y_i < Y_j) - \frac{1}{4} \\ &= \frac{1}{n(n-1)} \sum_{i \neq j}^{n} I(X_{ik} < X_{jk}) I(H(Y_i) < H(Y_j)) - \frac{1}{4} \end{aligned} \tag{2.8}$$

3.4 Other marginal screening methods

- Tilting methods (Hall et al. 2009),
- Generallzed correlation screening (Hall and Miller 2009),
- Nonparametric screening (Fan et al.2011)
- Conditional Sure Independence Screening(Barut et al.2012)
- ...

4.1 Application: Financial Feature Screening for Stock Returns

Feature Screening for Network Autoregression Model

Statistica Sinica 2021

http://www3.stat.sinica.edu.tw/LatestART/SS-2018-0400_fp.pdf

1.模型建立

$$\boldsymbol{Y} = \rho \boldsymbol{W} \boldsymbol{Y} + \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

*类似空间联立自回归模型

ρ: 自相关系数

W: *权重*, $W_{ij} = \frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}}$

a_{ij}为节点间关系,*有联系赋值1*

$\varepsilon \sim (0, \sigma^2 I_n)$, ε 与X无关

2.筛选步骤

- 用(Y,WY)与X的多重相关系数作为排序依据 $\widehat{\mathrm{R}}_j^2 = \frac{\mathrm{X}_j^\top \{\widetilde{Y}(\widetilde{Y}^\top \widetilde{Y})^{-1} \widetilde{Y}^\top\} \mathrm{X}_j}{\mathrm{X}_j^\top \mathrm{X}_j},$
- 给定常数C_γ，选出模型 $\widehat{\mathcal{M}}^R = \left\{1 \le j \le p : \widehat{\mathrm{R}}_j^2 \ge c_\gamma\right\}.$

数据：

2014年在上海证券交易所和深圳证券交易所交易的487只A股股票

响应变量为同年对应年收益率

构建网络关系：共同股东

1. 收集各股头部前十位股东信息；
2. 对任意两只股票,若至少有1位共同股东,则标记为$a_{ij} = a_{ji}$ =1 (i≠j)
3. 记网络关系为邻接矩阵A,公司去年财务指标为协变量

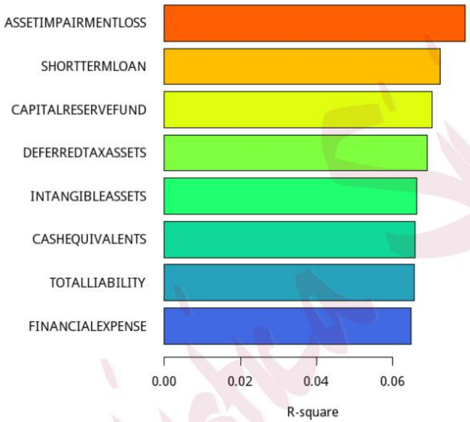


Figure 1: Covariates with top 8 $\widehat{\mathrm{R}}_j^2$. They are related to the asset (i.e., Asset Impairment Loss, Capital Reserve Fund, Deferred Tax Asset, Intangible Assets), liability (i.e., Short Term Loan, Total Liability), liquidity (i.e., Cash Equivalents), and Financial Expense of the firm.

4.1 Application: Financial Feature Screening for Stock Returns

3.评估准确度

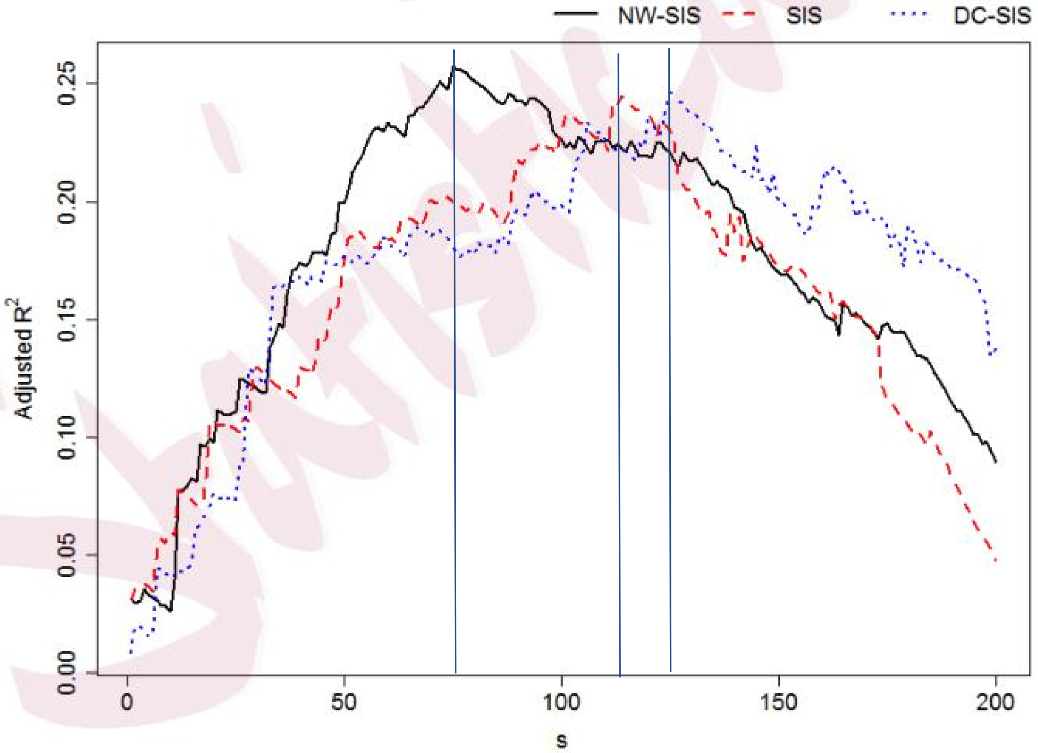
- 计算Y拟合值，计算线性回归后adjusted R^2

$$\widehat{Y} = (I - \widehat{\rho}_{\mathcal{M}} W)^{-1} \mathbb{X}_{\mathcal{M}} \widehat{\beta}_{\mathcal{M}}.$$

4.结论

- NW-SIS 在模型大小更小的时候，拟合优度相对表现较好

	NW-SIS	SIS	DC-SIS
adjusted R^2	0.259	0.247	0.248
Model Size	75	117	125



Summary

GOAL: Variable Screening

- **Proposed a two-scale learning framework:**
    - from p to d
    - from d to below size n for moderate-scale learning
  - Used marginal utilities based on marginal correlations $\widehat{corr}(x_j, y)$ *(sample correlation)*

    $$\widehat{M}_\gamma = \{1 \le j \le p : |\ \widehat{corr}(x_j, y)\ |\ is\ among\ the\ first\ [\gamma n]\ largest\ of\ all\}$$
  - SIS ideas can be incorporated into large-scale Bayesian estimation and inference
- **Proposed Sure Screening Property:**

$$\mathbb{P}\{\mathcal{M}_* \subset \widehat{\mathcal{M}}\} \to 1$$

- **Established theoretical foundation and examplar for subsequent research on this topic.**
  - Correlation-based /Model-based /Model-Free / Methods/...

感谢聆听！ 请大家批评指正！

THANK YOU FOR YOUR CRITICISM

presenter: Fei Yang 2021/04/30

Reference

- Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models, Fan(2011)
- Sure Independence Screening for Ultra-High Dimensional Feature Space, Fan(2008)
- Sure independence screening in generalized linear models with NP-dimensionality, Fan(2010)
- Additive Regression and Other Nonparametric Models, Stone(1985)
- A selective overview of feature screening for ultrahigh-dimensional data, Liu(2015)

2.1 NIS-Introduction

- ## Additive Model

$$Y = \sum_{j=1}^{p} m_j(x_j) + \varepsilon$$

$m_j(\cdot)$ $(j = 1,2, \ldots, p)$为非参数光滑函数

- ## NIS

  - consider independence learning by ranking the magnitude of marginal estimators, nonparametric marginal correlations, and the marginal residual sum of squares.

    - Sure Independence Screening
    - Iterative and Conditional Sure Independence Screening
    - Sure Independence Screening for Generalized Linear Models and Classification
    - Nonparametric and Robust Sure Independence Screening
    - Multivariate Sure Independence Screening and the Beyond