Sure Independence Screening for Ultrahigh Dimensional Feature Space

Jianqing Fan and Jinchi Lv

Journal of the Royal Statistical Society Series B. (2008)

Presenter: Jingjiang Peng

March 5, 2010

Outline

Introduction

- Consider the model selection problem in linear model

$$\mathrm{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

- AIC, BIC, best subset selection. NP-hard problem !
- LASSO: provide sparsity solution, model selection consistency: very strong conditions(Zhang and Yu(2006))
- SCAD: Oracle property, low dimension $\frac{p^3}{n} \to 0$ (Fan and Peng 2004)
- Adaptive LASSO: Oracle property, low dimension (Zou 2006)
- Dantzig selector: High dimension($p > n$)), Oracle property in the sense of Donoho and Johnstone. Need uniform uncertainty principle condition(UUP) (Candes and Tao 2007). Linear Programming is slow in ultrahigh dimension. p can not grow exponentially w.r.t n.

The challenges in high dimensional problems

$\mathrm{x} = (X_1, X_2, \ldots, X_p)^T$, and $\mathbf{\Sigma} = cov(\mathrm{x})$, $\mathrm{z} = \mathbf{\Sigma}^{-1/2}\mathrm{x}$. When p is larger than n, we will meet the following difficulties

- the matrix $\mathrm{X}^T\mathrm{X}$ is huge and singular. This causes trouble both in theory and computation
- the maximum spurious correlation between a covariate and the response can be very large, which makes the model selection difficult
- $\mathbf{\Sigma}$ may be singular or ill conditioned
- The minimum non-zero coefficients $|\beta_i|$ may decay close to noise level.
- The distribution of z may have heavy tails.

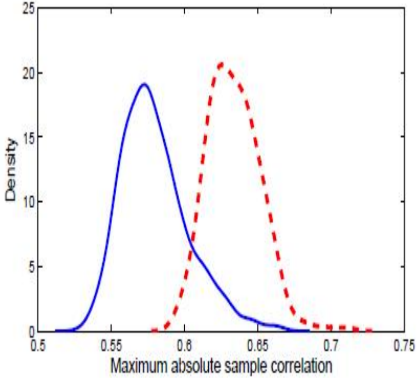

Figure 1: Distributions of the maximum absolute sample correlation coefficient when $n = 60, p = 1000$ (solid curve) and $n = 60, p = 5000$ (dashed curve).

Sure Independent Screening

- **Question**: Is there any model selection procedure that can effectively deal with ultrahigh dimensionality ($p = O(e^{n^\alpha})$) and keep the Oracle Property?

Sure Independent Screening

- **Question**: Is there any model selection procedure that can effectively deal with ultrahigh dimensionality ($p = O(e^{n^\alpha})$) and keep the Oracle Property?
- The answer: Sure Independency Screening (SIS), well in some sense!
- Let $\mathcal{M}_* = \{1 \le i \le p : \beta_i \ne 0\}$. The number of true non-zero coefficients $s = |\mathcal{M}_*|$, $\mathcal{M}_\gamma$ is the model selected by SIS with some parameter γ. $d = |\mathcal{M}_\gamma| = [\gamma n] < n$
- Main Result of SIS:
  Theorem 1: Under some regular conditions,
  $$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \qquad (2)$$
- SIS-SCAD or SIS-adaptive Lasso on $\mathcal{M}_\gamma$ can achieve Oracle Property

SIS: A correlation learning method

- Suppose X has been standardized The componentwise regression is

$$\mathrm{w} = \mathrm{X}^T \mathrm{y} \tag{3}$$

- SIS: For any given $\gamma \in (0, 1)$, sort the p componentwise magnitudes of the vector w in a decreasing order

$$\mathcal{M}_\gamma = \{1 \le i \le p : \ |w_i| \text{ is among the first } [\gamma n] \text{ largest of all}\} \tag{4}$$

- SIS selects $d = [\gamma n] < n$ parameters, and reduce the dimension less than n. SCAD, adaptive LASSO, Dantzig selector can applied to achieve good properties, if SIS satisfies sure screening property

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1 \tag{5}$$

- SCAD:

$$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathrm{x}_i'\boldsymbol{\beta})^2 + \sum_{i=1}^{d} p_\lambda(|\beta_j|)$$

  where $p_\lambda(|\beta_j|)$ is the SCAD penalty

- Adaptive Lasso:

$$min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathrm{x}_i'\boldsymbol{\beta})^2 + \lambda\sum_{i=1}^{d} w_j|\beta_j|$$

  where w_j is the adaptive weight. Usually, it is related to least square estimator

- Dantzig selector

$$min\|\zeta\|_1 \ \text{ subject to } \|\mathrm{X}'_{\mathcal{M}}r\|_\infty \le \lambda_d\sigma$$

  where $\lambda_d > 0$ and $r = \mathrm{y} - \mathrm{X}_{\mathcal{M}}\zeta$

Iteratively Thresholded Ridge Regression Screener (ITRRS)

- Consider the ridge regression

$$\mathrm{w}^{\lambda} = (\mathrm{X}^T\mathrm{X} + \lambda I_p)^{-1}\mathrm{X}^T\mathrm{y} \tag{6}$$

$$\mathrm{w}^{\lambda} \to \hat{\boldsymbol{\beta}}_{LS} \quad as \ \lambda \to 0$$

$$\lambda\mathrm{w}^{\lambda} \to \mathrm{w} \quad as \ \ \lambda \to \infty$$

- For any given $\delta \in (0, 1)$, sort the p componentwise magnitudes of the vector w^{λ} in a descending order, and define

$$\mathcal{M}^1_{\delta,\lambda} = \{1 \le i \le p : \ |w_i^{\lambda}| \text{ is among the first } [\delta p] \text{ largest of all}\} \tag{7}$$

- This procedure reduces the model by a factor $(1 - \delta)$. This procedure can be applied iteratively until the remaining variable is less than n

Steps of ITRRS

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

Main result of ITRRS:

Theorem 3: Under some regular conditions,

$$P(\mathcal{M}_* \subset \mathcal{M}_{\delta,\lambda}) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \tag{8}$$

Regularity Conditions

- *condition* 1. $p > n$ and $log(p) = O(n^{\xi})$ for some $\xi \in (0, 1 - 2\kappa)$
- *condition* 2. z has a spherically symmetric distribution. Let $Z = (z_1, \ldots, z_n)^T$, and there are some $c, c_1 > 1$ and $C_1 > 0$ such that

$$P\{\lambda_{max}(\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T) > c_1 \text{ or } \lambda_{min}(\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T) < 1/c_1\} \leq exp(-C_1 n) \tag{9}$$

  holds for any $n \times \tilde{p}$ submatrix $\tilde{Z}$ of Z with $cn < \tilde{p} \leq p$. Also $\epsilon \sim N(0, \sigma^2)$
- *condition* 3. var(Y)=$O(1)$ and for some $\kappa \geq 0$ and $c_1, c_3 > 0$

$$min_{i \in \mathcal{M}_*} |\beta_j| \geq \frac{c_2}{n^{\kappa}} \;\; and \;\; min_{i \in \mathcal{M}_*} |cov(\beta_i^{-1} Y, X_i)| \geq c_3$$

- *condition* 4. There are some $\tau \geq 0$ and $c_4 > 0$ such

$$\lambda_{max}(\mathbf{\Sigma}) \leq c_4 n^{\tau} \tag{10}$$

Discussion about Regularity Conditions

- The main part of condition 2 means that the n non-zero singular value of the $n \times \tilde{p}$ matrix $\tilde{Z}$ are in the same order, which is reasonable. Because as $\tilde{p} \to \infty$, $\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T \to I_n$ by random matrix theory. This condition can be shared by a wide class of distribution.
- The first part of condition 3 tells us that the smallest absolute value of non-zero coefficients can be distinguished from noise. The second part rules our the situation in which an important variable is marginally uncorrelated with Y, but jointly correlated with Y.
- Although condition 4 allows the largest eigenvalue of $\mathbf{\Sigma}$ to diverge as n grows. We will see in the later theorem that τ must be a small number less than 1

Theorem 1

Thereom 1(accuracy of SIS): Under condition 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 $C > 0$

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \tag{11}$$

- let $\mathcal{O}(p)$ denote the orthogonal group, that is for any matrix $A_{p\times p} \in \mathcal{O}(p)$, $A^T A = 0$. By the condition 2, the distribution of z is invariant under $\mathcal{O}(p)$. Let $S^{q-1} = \{\mathrm{x} \in \mathbb{R}^q : \|\mathrm{x}\| = 1\}$ be the q dimensional unit ball.
- Let $\mu_1^{1/2}, \ldots, \mu_n^{1/2}$ be the singular value of Z, by SVD

$$Z_{n\times p} = V_{n\times n} D_{n\times p} U_{p\times p}$$

where $V \in \mathcal{O}(n)$, $U \in \mathcal{O}(p)$, $D = (diag(\mu_1^{1/2}, \ldots, \mu_n^{1/2}), 0, \ldots, 0)$

Lemma 1

- $S = (Z^T Z)^+ Z^T Z = U^T diag(I_n, 0) U$, where $(Z^T Z)^+$ is the Moore-Penrose generalized inverse.
- From the SVD, $(I_n, 0)_{n\times p} U = diag(1/\mu_1^{1/2}, \dots, 1/\mu_n^{1/2}) V^T Z$
- By condition 2, $ZQ =^d Z$ for any $Q \in \mathcal{O}(p)$
- Given V and $(\mu_1, \dots, \mu_n)^T$, the conditional distribution of $(I_n, 0)U$ is invariant under $\mathcal{O}(p)$

Lemma 1: $(I_n, 0)U =^d (I_n, 0)\bar{U}$ and $(\mu_1, \dots, \mu_n)^T$ is independent of $(I_n, 0)U$, where $\bar{U}$ is uniformly distributed on the orthogonal group $\mathcal{O}(p)$ and $\mu_1, \dots, \mu_n$ are n eigenvalues of ZZ^T

Lemma

Lemma 2: $< Se_1, e_1 > =^d \frac{\chi_n^2}{\chi_n^2 + \chi_{p-n}^2}$
Proof: $S =^d U^T diag(I_n, 0) U$ where U is uniformly distributed on $\mathcal{O}(p)$.
Ue_1 is a random vector uniformly distributed on the unit ball S^{p-1}.
Let $\mathrm{W} = (W_1, \ldots, W_p)^T \sim N(0, I_p)$ then $Ue_1 =^d \frac{\mathrm{W}}{\|\mathrm{W}\|}$, and

$$< Se_1, e_1 > = (Ue_1)^T diag(I_n, 0) Ue_1 =^d \frac{W_1^2 + \ldots W_n^2}{W_1^2 + \cdots + W_p^2}$$

Lemma 2 says $< Se_1, e_1 >$ is a beta distribution. By the property of Beta distribution we have
Lemma 4: For any $C > 0$, there are $0 < c_1 < 1 < c_2$, such that

$$P(< Se_1, e_1 > < c_1 \frac{n}{p} \text{ or } > c_2 \frac{n}{p}) \leq 4 \exp(-Cn) \qquad (12)$$

Lemma 5. Let $Se_1 = (V_1, V_2, \ldots, V_p)^T$, then, given $V_1 = v$, the random vector $(V_2, \ldots, V_p)^T$ is uniformly distributed on the sphere $S^{p-2}(\sqrt{v - v^2})$. Moreover, for any $C > 0$, there are some $c > 1$ such that

$$P(|V_2| > cn^{1/2}p^{-1}|W|) \leq 3\exp(-Cn) \tag{13}$$

where $W \sim N(0, 1)$

Main Idea of proof: Let $\mathrm{V} = (V_1, \ldots, V_p)^T$. For $Q \in \mathcal{O}(p-1)$, define $\tilde{Q} = diag(1, Q) \in \mathcal{O}(p)$, then

$$\tilde{Q}\mathrm{V} =^d (U\tilde{Q}^T)^T diag(I_n, 0)(U\tilde{Q}^T)\tilde{Q}e_1 =^d U^T diag(I_n, 0)Ue_1 =^d V$$

Similar to Lemma 2, conditional on V_1

$$V_2 =^d \sqrt{V_1 - V_1^2}\frac{W_1}{\sqrt{W_1^2 + \cdots + W_{p-1}^2}}$$

Where $W_1, \ldots, W_{p-1}$ are i.i.d standard normal.

Proof of theorem 1

Step 1: Let $\delta \in (0,1)$, define the submodel

$$\tilde{\mathcal{M}}_\delta^1 = \{1 \le i \le p : \ |w_i| \text{ is among the first } [\delta p] \text{ largest of all}\} \tag{14}$$

Show that

$$P(\mathcal{M}_* \subset \tilde{\mathcal{M}}_\delta^1) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \tag{15}$$

$X = Z\mathbf{\Sigma}^{1/2}$, and

$$X^T X = p\mathbf{\Sigma}^{1/2}\tilde{U}^T diag(\mu_1, \dots, \mu_n)\tilde{U}\mathbf{\Sigma}^{1/2}$$

Here $\mu_1, \dots \mu_n$ are n eigenvalue of $p^{-1}ZZ^T$, $\tilde{U} = (I_n, 0)_{n\times p}U$

$$\mathrm{w} = X^T X\boldsymbol{\beta} + X^T\boldsymbol{\epsilon} = \boldsymbol{\xi} + \boldsymbol{\eta}$$

Step 1.1: Deal with $\boldsymbol{\xi}$
Step 1.1.1: Bounding $\|\boldsymbol{\xi}\|$ from above:

$$P\{\|\boldsymbol{\xi}\|^2 > O(n^{1+\tau}p)\} \leq O(\exp(-Cn))$$

First

$$\|\boldsymbol{\xi}\|^2 \leq p^2\lambda_{max}(\boldsymbol{\Sigma})\lambda_{max}(p^{-1}ZZ^T)^2\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{1/2}\tilde{U}^T\tilde{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}$$

Let $Q \in \mathcal{O}(p)$ such that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|Qe_1$,then

$$\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{1/2}\tilde{U}^T\tilde{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} =^d \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 < Se_1, e_1 >$$

By Lemma 4 and $|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 = \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \leq var(Y) = O(1)$

$$P\{\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{1/2}\tilde{U}^T\tilde{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} > O(\frac{n}{p})\} \leq O(\exp(-Cn))$$

Finally note $\lambda_{max}(\boldsymbol{\Sigma}) = O(n^\tau)$ and $P\{\lambda_{max}(p^{-1}ZZ^T) > c_1\} \leq \exp(-C_1n)$

Step 1.1.2: Bounding $\|\xi_i\|$, $i \in \mathcal{M}_*$ from above:

$$P(|\xi_i| < cn^{1-\kappa}) \leq O[\exp\{-Cn^{1-2\kappa}/log(n)\}] \quad \imath \in \mathcal{M}_* \tag{16}$$

Step 1.2 Deal with $\boldsymbol{\eta}$
Step 1.2.1: Bounding $\|\boldsymbol{\eta}\|$ from above:

$$P\{\|\boldsymbol{\eta}\|^2 > O(n^{1+\tau}p)\} \leq O(\exp(-Cn)) \tag{17}$$

Step 1.2.2: Bounding $|\eta_i|$ from above:

$$P\{max_i|\eta_i| > o(n^{1-\kappa})\} \leq O[\exp\{-Cn^{1-2\kappa}/log(n)\}] \tag{18}$$

Setp 1.3: Combine the result in 1.1 and 1.2, we have

$$P(min_{i\in\mathcal{M}_*}|w_i| < c_1 n^{1-\kappa} \text{ or } \|\mathrm{w}\|^2 > c_2 n^{1+\tau}p) \leq O[s\exp\{-Cn^{1-2\kappa}/log(n)\}] \tag{19}$$

The above equation imply that for some $c > 0$

$$\#\{1 \leq k \leq p : |w_k| \geq min_{i \in \mathcal{M}_*} |w_i|\} \leq c \frac{n^{1+\tau} p}{(n^{1-\kappa})^2} = \frac{cp}{n^{1-2\kappa-\tau}} \qquad (20)$$

If we choose δ such that $\delta n^{1-2\kappa-\tau} \to \infty$ then

$$P(\mathcal{M}_* \subset \tilde{\mathcal{M}}_\delta^1) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \qquad (21)$$

holds for some constant $C > 0$

Step 2: Fix arbitrary $r \in (0,1)$ and choose the shrinking factor δ of the form $(n/p)^{1/(k-r)}$ for some integer $k \geq 1$.

- Carry out procedure (14) and obtain a submodel $\tilde{\mathcal{M}}_\delta^1$ with size $[\delta p]$
- Apply the similar procedure to model $\tilde{\mathcal{M}}_\delta^1$ to obtain a submodel $\tilde{\mathcal{M}}_\delta^2 \subset \tilde{\mathcal{M}}_\delta^1$ with $[\delta^2 p]$, and go on
- Finally obtain a submodel $\tilde{\mathcal{M}}_\delta = \tilde{\mathcal{M}}_\delta^k$ with size $d = [\delta^k p] = [\delta^r n] < n$, where $[\delta^{k-1} p] = [\delta^{r-1} n] > n$

It is easy to see $\tilde{\mathcal{M}}_\delta = \mathcal{M}_\gamma$ where $\gamma = \delta^r < 1$

How to choose δ?

For fixed $\theta_1 \in (0, 1-2\kappa-\tau)$ and pick some $r < 1$ very close to 1 such that $\theta_0 = \frac{\theta_1}{r} < 1 - 2\kappa - \tau$. Choose δ such that

$$\delta n^{1-2\kappa-\tau} \to \infty \quad \text{and} \quad \delta n^{\theta_0} \to 0$$

The corresponding γ is

$$\gamma n^{r(1-2\kappa-\tau)} \to \infty \quad \text{and} \quad \gamma n^{\theta_1} \to 0$$

Final Step to Prove Theorem 1

Based on the above steps

$$P(\mathcal{M}_* \subset \tilde{\mathcal{M}}_\delta^i | \mathcal{M}_* \subset \tilde{\mathcal{M}}_\delta^{i-1}) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \tag{22}$$

Then

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O(kexp\{-Cn^{1-2\kappa}/log(n)\}) \tag{23}$$

Note by the requirement of δ, $k = O\{log(p)/log(n)\}$, which is of order $O(n^\xi/log(n))$. So

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \tag{24}$$

The condition of γ holds for $\gamma \sim cn^{-\theta}$ with $\theta < 1 - 2\kappa - \tau$

Theorem 2 and Theorem 3

- *Theorem* 2: (Asymptotic sure screening) Under condition 1-4, if $2\kappa + \tau < 1$, $\lambda(p^{3/2}n)^{-1} \to \infty$ and $\delta n^{1-2\kappa-\tau} \to \infty$, the we have for some $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}^1_{\delta,\gamma}) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \tag{25}$$

- *Theorem* 3:(Accuracy of ITRRS) Let the assumptions of theorem 2 be satisfied. If $\delta n^{\theta} \to \infty$ for come $\theta < 1 - 2\kappa - \tau$, then successive applications of ITRRS for k times results in a submodel $\mathcal{M}_{\delta,\lambda}$ with size $d = [\delta^k p] < n$ such that for some $C > 0$

$$P(\mathcal{M}_* \subset \mathcal{M}_{\delta,\gamma}) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \tag{26}$$

- The proofs are similar to theorem 1

Oracle Property of SIS-SCAD

Theorem 5: if $d = o(n^{1/3})$ and the assumptions of theorem in Fan and Peng (2004) are satisfied, then, with probability tending to 1, the SIS-SCAD estimator $\hat{\beta}_{SCAD}$ satisfies

- $\hat{\beta}_i = 0$ for any $i \notin \mathcal{M}_*$
- the components of $\hat{\beta}_{SCAD}$ in $\mathcal{M}_*$ perform as well as if the true model $\mathcal{M}_*$ were known

A Simulation Example

- Two models with $(n, p) = (200, 1000)$ and $(n, p) = (800, 20000)$. The sizes s of the true models are 8 and 18.
- The non-zero coefficients are randomly chosen as follows. Let $a = 4log(n)/n^{1/2}$ and $5log(n)/n^{1/2}$ for two different models, pick non-zero coefficients of the form $(-1)^u(a + |z|)$ for each model, where $u \sim Bernoulli(0.4)$ and $z \sim N(0, 1)$
- The l_2 norms $\|\beta\|$ of the two simulated models are set 6.795 and 8.908
- These settings are not trivial since there is non-negligible sample correlation between the predictors

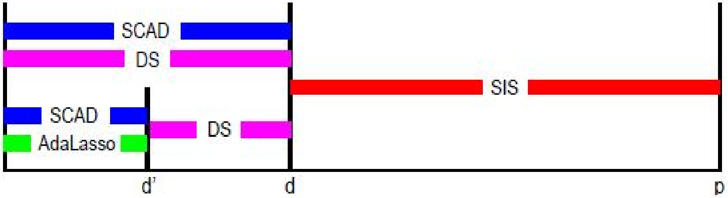

Figure 2: Methods of model selection with ultra high dimensionality.

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

Discussion

- Randomized design
- Are the regularization conditions reasonable?
- Correlation screening for Linear model. Are there any other screening methods for more general models?
- What is the relation between the correlation screening and multiple comparison?
- How to choose the tuning parameter γ, λ and δ?
- Σ may become singular when p is really large. Z is not well defined in this case.