Variable Selection in High Dimension

Jingjiang Peng (Jack)

Department of Statistics

April 21, 2010

Outline

1. Introduction
2. $p < n$ Case
   - SCAD
   - Adaptive LASSO
3. $p > n$ Case
   - LASSO
   - Dantzig Selector
   - Sure Independent Screening
4. Numerical Simulation

Introduction

- Consider the variable selection problem in linear model

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

where X is a $n \times p$ matrix. We are interested in $p > n$ case.

- Suppose the true $\boldsymbol{\beta}$ is $\boldsymbol{\beta}_0$ with support $\mathcal{A}$
- The aim of model selection is to identify $\mathcal{A}$ as close as possible

Criteria for Good Variable Selection Procedure

- Suppose that we have an estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ by n observations

Criteria for Good Variable Selection Procedure

- Suppose that we have an estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ by n observations
- **Model Selection Consistency:** $P(\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}) \to 1$ as $n \to \infty$
- Model selection consistency is stronger than the ordinary consistency of parameter estimator.

Criteria for Good Variable Selection Procedure

- Suppose that we have an estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ by n observations
- **Model Selection Consistency:** $P(\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}) \rightarrow 1$ as $n \rightarrow \infty$
- Model selection consistency is stronger than the ordinary consistency of parameter estimator.
- **Sign Consistency:** $P(sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}_0)) \rightarrow 1$ as $n \rightarrow \infty$

Criteria for Good Variable Selection Procedure

- Suppose that we have an estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ by n observations
- **Model Selection Consistency:** $P(\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}) \rightarrow 1$ as $n \rightarrow \infty$
- Model selection consistency is stronger than the ordinary consistency of parameter estimator.
- **Sign Consistency:** $P(sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}_0)) \rightarrow 1$ as $n \rightarrow \infty$
- **Oracle Property:**
  - Selection consistency: $P(\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}) \rightarrow 1$
  - Asymptotic normality: $\sqrt{n}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A},0}) \rightarrow N(0, C_{\mathcal{A},\mathcal{A}})$, where $\frac{1}{n}X'X \rightarrow C$

  The oracle property says that our estimator has the same efficiency as estimator of $\boldsymbol{\beta}_{\mathcal{A}}$ based on the submodel with $\boldsymbol{\beta}_{\mathcal{A}^c} = 0$ known in advance

Variable Selection when $p < n$

- AIC, BIC, subset selection: Combinatoric, NP hard, computational intensive when p is large

Variable Selection when $p < n$

- AIC, BIC, subset selection: Combinatoric, NP hard, computational intensive when p is large
- LASSO: $min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\beta)^2 + \lambda\sum_{i=1}^{d}|\beta_j|$

Variable Selection when $p < n$

- AIC, BIC, subset selection: Combinatoric, NP hard, computational intensive when p is large
- LASSO: $min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\beta)^2 + \lambda\sum_{i=1}^{d}|\beta_j|$
- LASSO is not model selection consistent for general design X

Variable Selection when $p < n$

- AIC, BIC, subset selection: Combinatoric, NP hard, computational intensive when p is large
- LASSO: $min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \lambda\sum_{i=1}^{d}|\beta_j|$
- LASSO is not model selection consistent for general design X
- Bridge: $min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \lambda\sum_{i=1}^{p}|\beta_j|^{\gamma}$ where $0 < \gamma < 1$

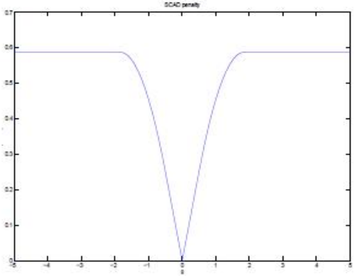
Variable Selection when $p < n$

- AIC, BIC, subset selection: Combinatoric, NP hard, computational intensive when p is large
- LASSO: $min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \lambda\sum_{i=1}^{d}|\beta_j|$
- LASSO is not model selection consistent for general design X
- Bridge: $min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \lambda\sum_{i=1}^{p}|\beta_j|^{\gamma}$ where $0 < \gamma < 1$
- SCAD: $min\ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \sum_{i=1}^{p}p_{\lambda}(|\beta_j|)$

Variable Selection when $p < n$

- AIC, BIC, subset selection: Combinatoric, NP hard, computational intensive when p is large
- LASSO: $min \ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \lambda \sum_{i=1}^{d} |\beta_j|$
- LASSO is not model selection consistent for general design X
- Bridge: $min \ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \lambda \sum_{i=1}^{p} |\beta_j|^\gamma$ where $0 < \gamma < 1$
- SCAD: $min \ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \sum_{i=1}^{p} p_\lambda(|\beta_j|)$
- Adaptive Lasso: $min \ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{i=1}^{p} w_j|\beta_j|$

SCAD



$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\} \tag{2}$$

for some $a > 2$ and $\theta > 0$. It is a quadratic spline function with two knots at λ and $a\lambda$.

SCAD

Theorem

If $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ *and* $liminf_{n\to\infty} liminf_{\theta\to 0^+} \frac{p'_{\lambda_n}(\theta)}{\lambda_n} > 0$ *then there exists a local minimizer such that*

- *Selection consistency:* $P(\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}) \to 1$
- *Asymptotic normality:* $\sqrt{n}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A},0}) \to N(0, C_{\mathcal{A},\mathcal{A}})$

One shortcoming of SCAD is that it is not convex.

Adaptive LASSO

$$min \frac{1}{2n}\sum_{i=1}^{n}(Y_i - x_i'\beta)^2 + \lambda_n \sum_{i=1}^{p} w_j|\beta_j|$$

The weights is chosen by $w = 1/|\hat{\beta}|^\gamma$ where $\hat{\beta}$ is the OLS

Theorem

if $\sqrt{n}\lambda \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Then the adaptive lasso estimates must satisfy the following:

- *Selection consistency:* $P(\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}) \to 1$
- *Asymptotic normality:* $\sqrt{n}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A},0}) \to N(0, C_{\mathcal{A},\mathcal{A}})$

Adaptive LASSO is convex. It can be efficiently solved by LAR algorithm

Variable Selection in High Dimension – Overview

- AIC, BIC, best subset selection fails

Variable Selection in High Dimension – Overview

- AIC, BIC, best subset selection fails
- LASSO: still provide sparsity solution, model selection consistency: Irrepresentable conditions

Variable Selection in High Dimension – Overview

- AIC, BIC, best subset selection fails
- LASSO: still provide sparsity solution, model selection consistency: Irrepresentable conditions
- Direct SCAD, adaptive LASSO fails. We need to do some modification.

Variable Selection in High Dimension – Overview

- AIC, BIC, best subset selection fails
- LASSO: still provide sparsity solution, model selection consistency: Irrepresentable conditions
- Direct SCAD, adaptive LASSO fails. We need to do some modification.
- Dantzig selector: (Candes and Tao 2007)

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

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

Variable Selection in High Dimension – Overview

- AIC, BIC, best subset selection fails
- LASSO: still provide sparsity solution, model selection consistency: Irrepresentable conditions
- Direct SCAD, adaptive LASSO fails. We need to do some modification.
- Dantzig selector: (Candes and Tao 2007)

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

  where $\lambda_d > 0$ and $r = y - X_{\mathcal{M}}\zeta$
- Sure Independence Screen: Two step procedure, first reduce the dimension by screening than do model selection.

LASSO in High Dimensions

Definition

Irrepresentable Condition: There exists a positive constant vector ξ such that

$$|C_{\mathcal{A},\mathcal{A}^c}(C_{\mathcal{A},\mathcal{A}})^{-1} sign(\beta_{\mathcal{A},0})| \leq 1 - \xi$$

Theorem

Under some technical regularity conditions, Irrepresentable Condition imples that LASSO sign consistency for $p_n = o(n^{ck})$. *for any* λ_n *satisfies* $\frac{\lambda_n}{\sqrt{n}} = o(n^{c/2})$ *and* $\frac{1}{p_n}(\frac{\lambda_n}{\sqrt{n}})^{2k} \to \infty$

Dantzig Selector

- In noiseless case, under RIP, one could recover $\boldsymbol{\beta}$ exactly by solving

$$min \sum_{i=1}^{p} |\beta_j|, \text{ subject to } X\boldsymbol{\beta} = y$$

- When the measurement device is subject to some small amount of noise. Candes and Tao (2007) proposed following convex program

$$min \sum_{i=1}^{p} |\beta_j|, \text{ subject to } \|X * r\|_\infty \leq \lambda_p \sigma$$

  for some $\lambda_p > 0$, where $r = y - X\boldsymbol{\beta}$ is residual.
- This can be solved by linear programming
- DS and LASSO are highly related. In many cases they provide the same solution path.(James, Radchenko and Lv 2009)

Good Propertis of DS

Theorem

Suppose β_0 is any S-sparse vector such that $\delta_{2S} + \theta_{S,2S} < 1$, choose $\lambda_p = \sqrt{2log(p)}$, then with large probability,

$$\|\hat{\beta} - \beta_0\|^2 \leq C_1 log(p) S\sigma^2$$

Some limitation of DS:

- RIP is too strong for statistics. Only random design can satisfy it. No fixed design can achieve this property at my knowledge.

Good Propertis of DS

Theorem

Suppose β_0 *is any S-sparse vector such that* $\delta_{2S} + \theta_{S,2S} < 1$*, choose* $\lambda_p = \sqrt{2log(p)}$*, then with large probability,*

$$\|\hat{\beta} - \beta_0\|^2 \leq C_1 log(p) S \sigma^2$$

Some limitation of DS:

- RIP is too strong for statistics. Only random design can satisfy it. No fixed design can achieve this property at my knowledge.
- p still can not too large. if $p = o(e^n)$ then the above theorem is useless in some sense.

Sure Independent Screening

- Suppose X has been standardized The componentwise regression is

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

Sure Independent Screening

- 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

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

Sure Independent Screening

- Suppose X has been standardized The componentwise regression is

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

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

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

- 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

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

Relation to the Ridge Regression

Consider the ridge regression

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

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

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

Theorem

Under some regularity conditions, 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* $C > 0$

$$P(\mathcal{A} \subset \mathcal{A}_\gamma) = 1 - O(exp\{-Cn^{1-2\kappa}/log(n)\}) \tag{7}$$

Theorem

(SIS-DS) Assume that $\delta_{2s} + \theta_{s,2s} \leq t < 1$*, and choose* $\lambda_d = \sqrt{2log(d)}$*, then with large probability,we have*

$$\|\hat{\beta} - \beta_0\|^2 \leq C\sqrt{log(d)}s\sigma^2$$

The above theorem reduce the factor $log(p)$ to $log(d)$ with $d < n$

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

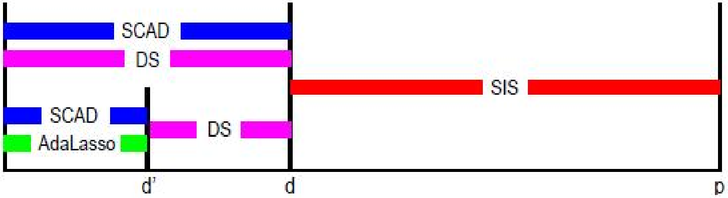


Figure 2: Methods of model selection with ultra high dimensionality.

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

Thank You!