Outline

1 Introduction

2 $p < n$ Case
   - SCAD
   - Adaptive LASSO

3 $p > n$ Case
   - LASSO
   - Dantzig Selector
   - Sure Independent Screening

4 Numerical Simulation
Consider the variable selection problem in linear model

\[ y = X\beta + \epsilon \]  

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

Suppose the true \( \beta \) is \( \beta_0 \) with support \( A \)

The aim of model selection is to identify \( A \) as close as possible
Criteria for Good Variable Selection Procedure

- Suppose that we have an estimator $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ by $n$ observations.
Criteria for Good Variable Selection Procedure

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

- **Model Selection Consistency:** \( P(\{j : \hat{\beta}_j \neq 0\} = A) \to 1 \) 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 $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ by $n$ observations.
- **Model Selection Consistency:** $P(\{j : \hat{\beta}_j \neq 0\} = A) \rightarrow 1$ as $n \rightarrow \infty$
- Model selection consistency is stronger than the ordinary consistency of parameter estimator.
- **Sign Consistency:** $P(\text{sign}(\hat{\beta}) = \text{sign}(\beta_0)) \rightarrow 1$ as $n \rightarrow \infty$
Criteria for Good Variable Selection Procedure

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

- **Model Selection Consistency:** $P(\{j : \hat{\beta}_j \neq 0\} = A) \to 1$ as $n \to \infty$

- Model selection consistency is stronger than the ordinary consistency of parameter estimator.

- **Sign Consistency:** $P(\text{sign}(\hat{\beta}) = \text{sign}(\beta_0)) \to 1$ as $n \to \infty$

- **Oracle Property:**
  - Selection consistency: $P(\{j : \hat{\beta}_j \neq 0\} = A) \to 1$
  - Asymptotic normality: $\sqrt{n}(\hat{\beta}_A - \beta_{A,0}) \to N(0, C_{A,A})$, where $\frac{1}{n}X'X \to C$

  The oracle property says that our estimator has the same efficiency as estimator of $\beta_A$ based on the submodel with $\beta_{A^c} = 0$ known in advance.
• 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: Combinatorial, 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: \[ \min \frac{1}{2n} \sum_{i=1}^{n} (Y_i - x'_i \beta)^2 + \lambda \sum_{i=1}^{d} |\beta_j| \]
- LASSO is not model selection consistent for general design $X$
- Bridge: \[ \min \frac{1}{2n} \sum_{i=1}^{n} (Y_i - x'_i \beta)^2 + \lambda \sum_{i=1}^{p} |\beta_j|^\gamma \text{ where } 0 < \gamma < 1 \]
Variable Selection when $p < n$

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

Bridge: $\min \frac{1}{2n} \sum_{i=1}^{n} (Y_i - x_i' \beta)^2 + \lambda \sum_{i=1}^{p} |\beta_j|^\gamma$ where $0 < \gamma < 1$

SCAD: $\min \frac{1}{2n} \sum_{i=1}^{n} (Y_i - x_i' \beta)^2 + \sum_{i=1}^{p} p_\lambda(\beta_j)$

Adaptive Lasso: $\min \frac{1}{2n} \sum_{i=1}^{n} (Y_i - x_i' \beta)^2 + \lambda_n \sum_{i=1}^{p} w_j |\beta_j|$
\[ p'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_{+}}{(a - 1)\lambda} I(\theta > \lambda) \right\} \]  

for some \( a > 2 \) and \( \theta > 0 \). It is a quadratic spline function with two knots at \( \lambda \) and \( a\lambda \).
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:** \( P(\{j : \hat{\beta}_j \neq 0\} = A) \to 1 \)
- **Asymptotic normality:** \( \sqrt{n}(\hat{\beta}_A - \beta_{A,0}) \to N(0, C_{A,A}) \)

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 estimate.

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

Adaptive LASSO is convex. It can be efficiently solved by LAR algorithm.
AIC, BIC, best subset selection fails
- AIC, BIC, best subset selection fails
- LASSO: still provide sparsity solution, model selection consistency: Irrepresentable conditions
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.

\[
\min_{\zeta} \|\zeta\|_1 \quad \text{subject to} \quad \|X'Mr\|_\infty \leq \lambda d \sigma
\]

where \(\lambda d > 0\) and \(r = y - XM\zeta\)
Sure Independence Screen: Two step procedure, first reduce the dimension by screening than do model selection.
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: (Candès and Tao 2007)

\[
\min \| \zeta \|_1 \quad \text{subject to} \quad \| X'_r \|_{\infty} \leq \lambda_d \sigma
\]

where \( \lambda_d > 0 \) and \( r = y - X_M \zeta \)
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)

\[
\min \| \zeta \|_1 \quad \text{subject to} \quad \| X'_M r \|_\infty \leq \lambda_d \sigma
\]

where \( \lambda_d > 0 \) and \( r = y - X_M \zeta \)

- Sure Independence Screen: Two step procedure, first reduce the dimension by screening than do model selection.
**Definition**

**Irrepresentable Condition:** There exists a positive constant vector $\xi$ such that

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

**Theorem**

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

• In noiseless case, under RIP, one could recover $\beta$ exactly by solving

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

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

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

for some $\lambda_p > 0$, where $r = y - X\beta$ 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 Properties of DS

**Theorem**

Suppose $\beta_0$ is any $S$-sparse vector such that $\delta_{2S} + \theta_{S,2S} < 1$, choose $\lambda_p = \sqrt{2\log(p)}$, 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 Properties of DS

Theorem

Suppose $\beta_0$ is any $S$-sparse vector such that $\delta_{2S} + \theta_{S,2S} < 1$, choose $\lambda_p = \sqrt{2 \log(p)}$, 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

\[ w = X^T y \]  \hspace{1cm} (3)
Suppose $X$ has been standardized. The componentwise regression is

$$w = X^T y$$  \hfill (3)$$

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

$$\mathcal{A}_\gamma = \{1 \leq i \leq p : |w_i| \text{ is among the first } \lfloor \gamma n \rfloor \text{ largest of all}\}$$  \hfill (4)$$
Suppose $X$ has been standardized. The componentwise regression is
\[ w = X^T y \]  \hspace{1cm} (3)

SIS: For any given $\gamma \in (0, 1)$, sort the $p$ componentwise magnitudes of the vector $w$ in a decreasing order
\[ \mathcal{A}_\gamma = \{1 \leq i \leq p : |w_i| \text{ is among the first } \lceil \gamma n \rceil \text{ largest of all} \} \]  \hspace{1cm} (4)

SIS selects $d = \lceil \gamma n \rceil < n$ 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
\[ P(\mathcal{A} \subset \mathcal{A}_\gamma) \to 1 \]  \hspace{1cm} (5)
Consider the ridge regression

\[ w^\lambda = (X^T X + \lambda I_p)^{-1}X^T y \]  

\[ w^\lambda \to \hat{\beta}_{LS} \quad \text{as} \quad \lambda \to 0 \]

\[ \lambda w^\lambda \to w \quad \text{as} \quad \lambda \to \infty \]
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 \( C > 0 \)

\[
P(A \subset A_\gamma) = 1 - O\left(\exp\{-Cn^{1-2\kappa}/\log(n)\}\right)
\tag{7}
\]

Theorem

(SIS-DS) Assume that \( \delta_{2s} + \theta_{s,2s} \leq t < 1 \), and choose \( \lambda_d = \sqrt{2\log(d)} \), 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 
  \(a = 4\log(n)/n^{1/2}\) and \(5\log(n)/n^{1/2}\) 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

<table>
<thead>
<tr>
<th>$p$</th>
<th>DS</th>
<th>Lasso</th>
<th>SIS-SCAD</th>
<th>SIS-DS</th>
<th>SIS-DS-SCAD</th>
<th>SIS-DS-AdaLasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>$10^3$</td>
<td>62.5</td>
<td>15</td>
<td>37</td>
<td>27</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>1.381</td>
<td>0.895</td>
<td>0.374</td>
<td>0.795</td>
<td>0.614</td>
<td>1.269</td>
</tr>
<tr>
<td>20000</td>
<td>—</td>
<td>—</td>
<td>37</td>
<td>119</td>
<td>60.5</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>—</td>
<td>0.288</td>
<td>0.732</td>
<td>0.372</td>
<td>1.014</td>
</tr>
</tbody>
</table>
Thank You!