Variable Selection Using the Dantzig Selector: Theoretical Properties and Extensions

Lee Dicker and Xihong Lin

Department of Biostatistics
School of Public Health
Harvard University

March, 2009
• Motivation
• Dantzig Selector (DZ)
• Large Sample Properties of Dantzig Selector
• Adaptive (Doubly Weighted) Dantzig Selector
• Relationship of Adaptive DS and Adaptive LASSO
• Simulation Studies
• Conclusions
The rapid advance of biotechnology yields massive high-throughput ”omics” data.

Examples include genomics, epigenomics, proteomics, metabolomics, ..., X-omics.

Such ”omics” technology has been applied rapidly to population-based studies to study interplay of gene and environment in causing human diseases.
Motivation: "Omics" Data in Population-Based Studies

- In genome-wide association studies, a million SNP markers are genotyped across the genome to study the association of common genetic variants and disease phenotypes (case/control status).

- The massive whole genome-wide sequencing data are rapidly available, e.g., the 1000 genome project, to study the effects of rare variants.
Variable selection is of significant interests in current biomedical ”omics” studies.

Common variable selection approaches are penalized likelihood based, e.g., LASSO, adaptive LASSO and SCAD.

The Dantzig selector (DS; Candes and Tao, 2007) can be viewed as an estimating equation based variable selection procedure and is particularly appealing for longitudinal data.

Little is known about the asymptotic properties of the DS.

Focus of this talk (Independent data):
(1) Large sample properties of the DS. (2) Propose Adaptive DS. (3) Connection with Adaptive Lasso.
Example: GWAS on Childhood Neuro-development in Mexico

- $n = 1000$ newborns
- 620,000 SNP markers (0,1,2) across the genome
- Outcome: Baylor score of neurodevelopment at 12 months.
- Interested in studying which genes are associated with early childhood neuro-development and gene-metal exposure interactions.
Model Setup

Model:

\[ y_i = X_i^T \beta + \epsilon_i, \]

where \( X_i = \) covariates \((p \times 1)\) and \( \beta = \) regression coefficients \((p \times 1)\) and \( E(\epsilon_i) = 0 \) and \( E(\epsilon_i^2) = \sigma^2 \).

Problem:

- Suppose the true set of non-zero \( \beta \)s: \( T^* = \{j : \beta_j \neq 0\} \).
- Goal: Identify \( T^* \) and estimate \( \beta_T = \{\beta_j\}_{j \in T^*} \).
Penalized Likelihoods and Lasso

- Penalized likelihood: Simultaneous model selection and estimation.
  - Maximize
    \[-\log(\text{likelihood}) + \text{sparseness penalty},\]

- Lasso
  \[
  \min_{\beta} \frac{1}{2} \| y - X\beta \|^2_2 + \lambda \| \beta \|_1.
  \] (lasso)
**Dantzig Selector**

- **Idea:** Rather than controlling the size of residuals, the Dantzig selector is based on the normal score equations and controls the correlation of residuals with $X$:

$$\text{minimize} \quad \|\beta\|_1$$

subject to \[ \|X'(y - X\beta)\|_\infty \leq \lambda. \]

(DS)


- The Dantzig selector and Lasso are closely related (Efron *et al.*, 2007; James, *et al.*, 2008).

- Little is known about the large sample properties of the DS.
Questions

- **Question(s) 1:**
  Is the Dantzig selector consistent for estimation? What is its asymptotic distribution?

- **Question 2:**
  Is the Dantzig selector consistent for model selection? *i.e.*, is
  
  \[
  \lim_{n \to \infty} P(\hat{T} = T^*) = 1?
  \]

- **Answers to Questions 1 and 2** depend heavily on choice of \( \lambda \).

- **When does the Dantzig selector have a unique solution?**
The DS feasible set: \( D = \{ \beta; \| X'(y - X\beta) \|_\infty \leq \lambda \} \)

If \( D \) is not parallel to the closed \( L_1 \)-unit ball, \( B_1 \), then the Dantzig selector has a unique solution.

If \( X_i \) are iid then the Dantzig selector has a unique solution with probability 1.
Asymptotic Properties of the Dantzig Selector

• If $C = \lim_{n \to \infty} n^{-1} X^T X$ is not parallel to the $L_1$-ball.

• Asymptotic Limit: If $\lambda/n \to c_0$ and $c_0 \in [0, \infty]$, then $\hat{\beta} \overset{P}{\to} \beta^*$, where $\beta^*$ is the true value of $\beta$ and $\beta_0$ solves

\[
\minimize ||\beta||_1
\]

subject to $||C(\beta^* - \beta)||_\infty \leq c_0$. 

March, 2009

Xihong Lin
Asymptotic Properties of the Dantzig Selector

• If $\lambda/\sqrt{n} \rightarrow c_1$ and $c_1 \in [0, \infty)$, then $\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{D} u_0$, where $u_0$ solves

$$\minimize \sum_{j \in T^*} \text{sgn}(\beta^*_j)u_j + \sum_{j \notin T^*} |u_j|$$

subject to

$$||C(v - u)||_\infty \leq c_1,$$

and $v \sim N(0, \sigma^2 C^{-1})$.

• Model Selection Inconsistency: There exist a large class of matrices $C$ for which the Dantzig selector is not consistent for model selection, regardless of $\lambda$. 
Adaptive (Doubly Weighted) Dantzig Selector

● Ideas:
  – When we suspect $\beta_j = 0$, relax the constraint on the $j$-th score equation and heavily penalize non-zero $\beta_j$.
  – When we suspect $\beta_j \neq 0$, Nearly solve the $j$-th scoring equation and only moderately penalize non-zero $\beta_j$.

● Adaptively Dantzig Selector (ADS)

\[
\begin{align*}
\text{minimize} & \quad \sum_{j=1}^{n} w_j |\beta_j| \\
\text{subject to} & \quad |X_j'(y - X\beta)| \leq w_j \lambda, \quad j = 1, \ldots, p.
\end{align*}
\]

(ADS)

e.g., $w_j = |\beta_j^{LS}|^{-\gamma}$ for some $\gamma > 0$. 

March, 2009
Xihong Lin
Adaptive Dantzig Selector

Dantzig Selector

\[ \hat{\beta} + C^{-1} B_\infty \]

\[ \| \hat{\beta} \|_{1, B_1} \]

Doubly Weighted Dantzig Selector

\[ \hat{\beta} + C^{-1} W B_\infty \]

\[ \| W \hat{\beta} \|_{1, W^{-1} B_1} \]

DWDS, \( n = \infty \)

\[ \hat{\beta} = \beta ^* \]

\[ \| W \hat{\beta} \|_{1, W^{-1} B_1} \]

March, 2009 Xihong Lin
Weighted Dantzig Selector: Asymptotics

- Suppose that $\sqrt{n w_j / \lambda} = O_P(1)$ if $j \in T^*$ and $\sqrt{n w_j / \lambda} \to \infty$ if $j \notin T^*$.

- Model Selection Consistency: The adaptive Dantzig selector (ADS) is consistent for model selection:

\[
\lim_{n \to \infty} P(\hat{T} = T^*) = 1,
\]

- Oracle Properties: The ADS estimators are asymptotically equivalent to the OLS estimator of $\beta^*$ based on the true model $T^*$:

\[
\sqrt{n} \left( \hat{\beta}_{T^*} - \beta^*_{T^*} \right) \overset{D}{\to} N(0, \sigma^2 C_{T^*,T^*}^{-1}).
\]
Adaptive Dantzig Selector and Adaptive Lasso

- Adaptive LASSO:

\[ \min_{\beta} \frac{1}{2} ||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|. \]  

(alasso)

- Adaptive DS and Adaptive LASSO have the same asymptotic properties.
Set $\beta^0 = W\beta$ and $X_0 = XW^{-1}$, we have

$$\begin{align*}
\text{minimize} & \quad ||\beta^0||_1 \\
\text{subject to} & \quad ||X_0'(y - X_0\beta^0)||_{\infty} \leq \lambda
\end{align*}$$

(ADS)

and

$$\min_{\beta^0} \frac{1}{2} ||y - X_0\beta^0||_2^2 + \lambda ||\beta^0||_1$$

(ALASSO)

This implies ADS and DS have the same computational cost and one can implement ADS using DASSO.
Estimation of the Toning Parameter $\lambda$

- Degrees of Freedom for the Danzig Selection:

$$\hat{df}(\lambda) = \text{trace}\{X_T (X'_E X_T)^{-1} X_E\} = |T|,$$

where $E = \{j; |X'_j \{y - X \hat{\beta}\}| = \lambda\}$ and let $T = \{j; \hat{\beta}_j \neq 0\}$.

- BIC for DS and ADS:

$$\min_{\beta} (n\sigma^2)^{-1} ||y - X \hat{\beta}||^2_2 + n^{-1} \log(n) \hat{df}(\lambda).$$

- If $w_j = |\hat{\beta}_j (\text{OLS})|^{-1}$ and $\lambda$ is chosen to minimize the BIC, then the ADS is
  (i) consistent for model selection and
  (ii) asymp. equiv to $\beta^*_{LS}$ based on the true model, $T^*$. 

March, 2009

Xihong Lin
• Compare DS, adaptive DS, LASSO and Adaptive LASSO

• For adaptive DS (ADS) and Adaptive LASSO: \( w_j = \left| \hat{\beta}_{j \text{OLS}} \right|^{-1} \)

• Implementation:
  – LARS (Efron, et al., 2004) for lasso and alasso.
  – DASSO (an extension of LARS; James, et al., 2008) for Dantzig selector and ADS.
  – Both LARS and DASSO efficiently obtain estimates for all values of \( \lambda \).

• \( \lambda \) chosen to minimize prediction error using validation data (Data validation (DV)) and BIC.
Simulation Settings

- Simulation settings:

  \[ \beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0) \in \mathbb{R}^8, \]

  \[ X = (x_{ij}), \quad x_{ij} \sim N(0, 1), \quad \text{corr}(x_{ij}, x_{ij'}) = 0.5^{|j-j'|}, \quad i \neq j, \]

  with independent rows.

- Noise levels \( \sigma = 1 \).

- Number of observations: \( n = 50 \).

- 500 runs.
### Simulation Results

<table>
<thead>
<tr>
<th>Tuning</th>
<th>Estimation</th>
<th>Sq. Error</th>
<th>Model Error</th>
<th>Model Size</th>
<th>$F_+$</th>
<th>$F_-$</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>DV</td>
<td>DS</td>
<td>0.17</td>
<td>0.12</td>
<td>5.69</td>
<td>2.69</td>
<td>0.00</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>ADS</td>
<td>0.11</td>
<td>0.08</td>
<td>3.91</td>
<td>0.91</td>
<td>0.00</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.17</td>
<td>0.12</td>
<td>5.70</td>
<td>2.70</td>
<td>0.00</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>ALASSO</td>
<td>0.11</td>
<td>0.09</td>
<td>4.01</td>
<td>1.01</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>BIC</td>
<td>DS</td>
<td>0.19</td>
<td>0.15</td>
<td>4.08</td>
<td>1.08</td>
<td>0.00</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>ADS</td>
<td>0.12</td>
<td>0.09</td>
<td>3.22</td>
<td>0.22</td>
<td>0.00</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.18</td>
<td>0.14</td>
<td>4.10</td>
<td>1.10</td>
<td>0.00</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.12</td>
<td>0.09</td>
<td>3.23</td>
<td>0.23</td>
<td>0.00</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Concluding Remarks

- Adaptive DS has advantages over the Dantzig selector: consistency in model selection and oracle properties, parallel adaptive LASSO.
- ADS outperforms DS in finite sample simulation studies.
- More complex extensions of the Dantzig selector are possible.

\[
\text{LASSO} \quad \longrightarrow \quad \text{PL with penalty } p_\lambda
\]

\[
\text{Dantzig selector} \quad \longrightarrow \quad p_\lambda\text{-Dantzig selector}
\]

- Implementation and theory are more difficult with \( p_\lambda \)-DS.
- Extensions to generalized linear models and longitudinal data using estimating equations are in progress.