On Model selection and assessment

*LASSO, LARS and DANTZIG for variable selection*

Nov.12, 2008

Supervisor: Yizeng Liang
Reporter: Hongdong Li
Main Contents

1. About Model selection and assessment
2. LASSO: Least Absolute Shrinkage and Selection Operator
3. LARS: Least Angle Regression
4. DS: the Dantzig Selector
5. Conclusion
What is learning from data?

- Extract important patterns and trends
- Understand “what the data says.”

Please see preface, The Elements of Statistical Learning
Diabetes Data from our lab. and Xiangya Hospital

Table 1 Part of the metabolites’ concentration of healthy people and type II diabetes patients.

<table>
<thead>
<tr>
<th>ID</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>x6</th>
<th>x7</th>
<th>...</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.1560</td>
<td>0.0282</td>
<td>0.0843</td>
<td>0.0149</td>
<td>0.0300</td>
<td>1.8066</td>
<td>0.0496</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.2544</td>
<td>0.0257</td>
<td>0.0053</td>
<td>0.0086</td>
<td>0.0020</td>
<td>1.0900</td>
<td>0.0092</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.7151</td>
<td>0.0375</td>
<td>0.0542</td>
<td>0.0275</td>
<td>0.0146</td>
<td>4.3166</td>
<td>0.0645</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1.0503</td>
<td>0.0558</td>
<td>0.0574</td>
<td>0.0923</td>
<td>0.0420</td>
<td>3.7168</td>
<td>0.0605</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1.1591</td>
<td>0.0531</td>
<td>0.0345</td>
<td>0.0280</td>
<td>0.0371</td>
<td>3.1334</td>
<td>0.0455</td>
<td>...</td>
<td>-1</td>
</tr>
<tr>
<td>6</td>
<td>0.5987</td>
<td>0.0320</td>
<td>0.0235</td>
<td>0.0306</td>
<td>0.0173</td>
<td>5.3095</td>
<td>0.0365</td>
<td>...</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>2.1433</td>
<td>0.0677</td>
<td>0.0067</td>
<td>0.0486</td>
<td>0.1351</td>
<td>1.8396</td>
<td>0.0309</td>
<td>...</td>
<td>-1</td>
</tr>
<tr>
<td>8</td>
<td>1.4336</td>
<td>0.0627</td>
<td>0.0167</td>
<td>0.0460</td>
<td>0.0256</td>
<td>2.4983</td>
<td>0.0358</td>
<td>...</td>
<td>-1</td>
</tr>
<tr>
<td>9</td>
<td>0.9227</td>
<td>0.0427</td>
<td>0.0089</td>
<td>0.0308</td>
<td>0.0322</td>
<td>3.7735</td>
<td>0.0507</td>
<td>...</td>
<td>-1</td>
</tr>
</tbody>
</table>
GC/MS data of essential oil of a TCM (Traditional Chinese Medicine)
## Diabetes data from literature

<table>
<thead>
<tr>
<th>Patient</th>
<th>AGE</th>
<th>SEX</th>
<th>BMI</th>
<th>BP</th>
<th>...</th>
<th>Serum Measurements</th>
<th>...</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54</td>
<td>1</td>
<td>24.2</td>
<td>106</td>
<td>181</td>
<td>121</td>
<td>44.0</td>
<td>3.21</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>1</td>
<td>33.9</td>
<td>92</td>
<td>278</td>
<td>196</td>
<td>44.5</td>
<td>5.42</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>1</td>
<td>32.4</td>
<td>78</td>
<td>237</td>
<td>154</td>
<td>48.5</td>
<td>4.00</td>
</tr>
<tr>
<td>4</td>
<td>57</td>
<td>2</td>
<td>25.7</td>
<td>83</td>
<td>181</td>
<td>103</td>
<td>67.5</td>
<td>2.00</td>
</tr>
<tr>
<td>5</td>
<td>55</td>
<td>1</td>
<td>27.7</td>
<td>96</td>
<td>195</td>
<td>120</td>
<td>59.5</td>
<td>2.00</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
<td>1</td>
<td>31.8</td>
<td>102</td>
<td>200</td>
<td>104</td>
<td>58.5</td>
<td>2.00</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>441</td>
<td>38</td>
<td>2</td>
<td>23.2</td>
<td>89</td>
<td>162</td>
<td>97</td>
<td>51.5</td>
<td>2.00</td>
</tr>
<tr>
<td>442</td>
<td>36</td>
<td>2</td>
<td>26.5</td>
<td>11</td>
<td>232</td>
<td>159</td>
<td>40.5</td>
<td>5.00</td>
</tr>
</tbody>
</table>

**Table 1.** Diabetes study. 442 diabetes patients were measured on 10 baseline variables. A prediction model was desired for the response variable, a measure of disease progression one year after baseline.

Corn data: X: 80 x 700  y: 80x1
Linear model: y = X\beta + e
Model building, selection and assessment

Inputs $X$

Learning Algorithm

Model $y = f(x)$

Performance

Model assessment

Model selection

Outputs $\hat{y}$

Measured $y$
Central South University

Discriminant analysis
- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)
- Partial least squares discriminant analysis (PLS-DA)
- Bayes classifier
- Artificial neural networks (ANN)
- Support vector machines (SVMs)

Regression analysis
- Partial least squares (PLS)
- Artificial neural networks (ANN)
- Support vector machines (SVMs)
What are model selection and model assessment?

- **Model selection**: Estimating the performance of different models in order to choose the (approximate) best one.

- **Model assessment**: Have chosen a final model, estimating its prediction error (generalization error) on new data.

*Page 195 and 196, The Elements of Statistical Learning*
Why is model selection and assessment necessary?

Figure 7.1: Behavior of test sample and training sample error as the model complexity is varied.

The figure is taken from Pg 194 of the book *The Elements of Statistical Learning* by Hastie, Tibshirani and Friedman.
An example on model complexity: Fit a curve to the 17 points.
Occam’s Razor Theory: simplest explanation that accounts for the data is best

Taken from P. 345 *Information Theory, Inference, and Learning Algorithms* David J.C. MacKay
What should we do?

- Variable/feature Selection
- Controlling the complexity of a model
- Outlier detection
Sequential selection methods

Forward selection

Stepwise selection

Back elimination
Using information criteria (IC)

- General IC (GIC): $\hat{\log L}(\theta / y) - a \frac{k}{2}$
- Akaike's IC (AIC): $\log L(\theta / y) - k$
- Bayesian IC (BIC): $\log L(\theta / y) - \log \left( \frac{n}{2 \pi} \right) \frac{k}{2}$
$C_p$ statistic

$$C_p = p + \frac{(s^2 - s_{full}^2)}{S_{full}^2} (n - p)$$

$p$ The number of variables under consideration

$s^2$ The Mean Squared Error of model under consideration

$s_{full}^2$ The Mean Squared Error of model including all variables
Based on evolutionary algorithm

Using

**GA** Genetic algorithm

**PSO** Particle swarm optimization

**DE** Differential evolution
Some methods developed by chemometricians

Uninformative variable elimination (UVE)

Moving window PLS

Randomization test PLS
Let's see some recently developed methods for model selection by statisticians
Four methods

1. Least squares regression
2. Lasso (1996)
Least squares regression

Data: \( X \quad y \)

Model: \( y = X \beta + e \)

\( \hat{y} = X \hat{\beta} \)

Minimize: \( L = \sum_{i=1}^{N} (y_i - X \beta)^2 \)

Gradient: \( \frac{\partial L}{\partial \beta} = X^t (y - X \beta) \)

\( \hat{y} = X(X^t X)^{-1} X^t y = Hy \)
Now let's consider:

LS coefficients for diabetes data (442 x 10, $\mathbf{X}$ and $\mathbf{y}$ are both centered)

$$\| \beta_{LS} \|_1 = 3460.00 \quad \| \beta_{LS} \|_2 = 1377.85$$

$$\| \mathbf{y} - \mathbf{X} \beta_{LS} \|_2 = 1124.27 \quad \| \mathbf{X}^t (\mathbf{y} - \mathbf{X} \beta_{LS}) \|_\infty = 7.82 \times 10^{-13}$$
Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY

We propose a new method for estimation in linear models. The ‘lasso’ minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Johnstone. The lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.

Keywords: QUADRATIC PROGRAMMING; REGRESSION; SHRINKAGE; SUBSET SELECTION
2. THE LASSO

2.1. Definition

Suppose that we have data \((x^i, y_i), i = 1, 2, \ldots, N,\) where \(x^i = (x_{i1}, \ldots, x_{ip})^T\) are the predictor variables and \(y_i\) are the responses. As in the usual regression set-up, we assume either that the observations are independent or that the \(y_i\)s are conditionally independent given the \(x_{ij}\)s. We assume that the \(x_{ij}\)s are standardized so that \(\sum x_{ij}^2/N = 0, \sum x_{ij}^2/N = 1.\)

Letting \(\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T,\) the lasso estimate \((\hat{\alpha}, \hat{\beta})\) is defined by

\[
(\hat{\alpha}, \hat{\beta}) = \arg \min \left\{ \sum_{i=1}^N \left( y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to } \sum_j |\beta_j| \leq t. \quad (1)
\]

Here \(t \geq 0\) is a tuning parameter. Now, for all \(t,\) the solution for \(\alpha\) is \(\hat{\alpha} = \bar{y}.\) We can assume without loss of generality that \(\bar{y} = 0\) and hence omit \(\alpha.\)

Computation of the solution to equation (1) is a quadratic programming problem with linear inequality constraints. We describe some efficient and stable algorithms for this problem in Section 6.
L1 norm constraint

\[ \beta_2 \]

\[ \beta_1 \]

Green < Pink < Blue
Geometry of Lasso

Projection

\[ L(t) = \| y - \hat{y} \|_2 \]

Lasso constraint

\[ c = y - \hat{y} \]
Lasso Coefficient path for each variable (diabetes data)

\[ t = \sum |\hat{\beta}_j| \rightarrow \]

\[ \Box \beta_{LS} \Box = 3460.00 \]
Least Angle Regression

Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani
Statistics Department, Stanford University

February 27, 2002

Abstract

The purpose of model selection algorithms such as All Subsets, Forward Selection, and Backward Elimination is to choose a linear model on the basis of the same set of data to which the model will be applied. Typically we have available a large collection of possible covariates from which we hope to select a parsimonious set for the efficient prediction of a response variable. Least Angle Regression ("LARS"), a new model selection algorithm, is a useful and less greedy version of traditional forward selection methods. Three main properties are derived. (1) A simple modification of the LARS algorithm implements the Lasso, an attractive version of Ordinary Least Squares that constrains
Central South University

Born in 1938

Awarded 2005 National Medal of Science and Technology
Some elements

- Equiangular vector of the active variables
- Only need $P$ iterations if $p$ variables
- A 'democratic' method, not too greedy
Lars algorithm

Initialize: 0. \( X, y, \mu_0 = 0, \text{Var} = [ ] \)

Loop start 2. \( c = X'(y - \mu_0) \)

3. \( j = \arg\max_i (|c_i|) \text{ Var} = [\text{Var } j] \)

4. \( s_j = \text{sgn}(c(\text{Var})) \)

5. \( X_{\text{active}} = [\ldots s_jx_j \ldots], j \in \text{Var} \)

6. \( u, X'_{\text{active}}u = \text{const tan t}*1 \)

7. \( \mu_1 = \mu_0 + \gamma u \)

Loop end 8. \( \mu_0 = \mu_1 \)
Correlative coefficients with equiangular vector

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1851</td>
<td>0.2681</td>
<td>0.3410</td>
<td>0.2962</td>
<td>0.1081</td>
<td>0.1361</td>
<td>0.0499</td>
<td>0.0248</td>
<td>0.0430</td>
</tr>
<tr>
<td>2</td>
<td>0.0822</td>
<td>0.1400</td>
<td>0.2078</td>
<td>0.3324</td>
<td>-0.4664</td>
<td>-0.4537</td>
<td>-0.3602</td>
<td>-0.2685</td>
<td>-0.0697</td>
</tr>
<tr>
<td>3</td>
<td>1.0000</td>
<td>0.8503</td>
<td>0.7794</td>
<td>0.7169</td>
<td>0.4664</td>
<td>0.4537</td>
<td>0.3602</td>
<td>0.2685</td>
<td>0.0697</td>
</tr>
<tr>
<td>4</td>
<td>0.3954</td>
<td>0.4639</td>
<td>0.7794</td>
<td>0.7169</td>
<td>0.4664</td>
<td>0.4537</td>
<td>0.3602</td>
<td>0.2685</td>
<td>0.0697</td>
</tr>
<tr>
<td>5</td>
<td>0.2498</td>
<td>0.4500</td>
<td>0.4279</td>
<td>0.2790</td>
<td>0.1352</td>
<td>0.1554</td>
<td>-0.3602</td>
<td>-0.2685</td>
<td>-0.0697</td>
</tr>
<tr>
<td>6</td>
<td>0.2612</td>
<td>0.3408</td>
<td>0.3248</td>
<td>0.3174</td>
<td>0.1487</td>
<td>0.1726</td>
<td>-0.4174</td>
<td>-0.3258</td>
<td>0.0697</td>
</tr>
<tr>
<td>7</td>
<td>-0.3668</td>
<td>-0.4501</td>
<td>-0.3990</td>
<td>-0.7169</td>
<td>-0.4664</td>
<td>-0.4537</td>
<td>-0.3602</td>
<td>-0.2685</td>
<td>-0.0697</td>
</tr>
<tr>
<td>8</td>
<td>0.4138</td>
<td>0.6066</td>
<td>0.5455</td>
<td>0.7035</td>
<td>0.4104</td>
<td>0.4182</td>
<td>0.0574</td>
<td>0.2685</td>
<td>0.0697</td>
</tr>
<tr>
<td>9</td>
<td>0.4462</td>
<td>0.8503</td>
<td>0.7794</td>
<td>0.7169</td>
<td>0.4664</td>
<td>0.4537</td>
<td>0.3602</td>
<td>0.2685</td>
<td>0.0697</td>
</tr>
<tr>
<td>10</td>
<td>0.3887</td>
<td>0.5018</td>
<td>0.5311</td>
<td>0.5113</td>
<td>0.2673</td>
<td>0.4537</td>
<td>0.3602</td>
<td>0.2685</td>
<td>0.0697</td>
</tr>
</tbody>
</table>
Lars coefficient path for each variable (diabetes data)
Lars vs Lasso (real vs broken)
THE DANTZIG SELECTOR: STATISTICAL ESTIMATION WHEN $p$
IS MUCH LARGER THAN $n^1$

BY EMMANUEL CANDÈS$^2$ AND TERENCE TAO$^3$

California Institute of Technology and University of California, Los Angeles

In many important statistical applications, the number of variables or parameters $p$ is much larger than the number of observations $n$. Suppose then that we have observations $y = X\beta + z$, where $\beta \in \mathbb{R}^p$ is a parameter vector of interest, $X$ is a data matrix with possibly far fewer rows than columns, $n \ll p$, and the $z_i$'s are i.i.d. $N(0, \sigma^2)$. Is it possible to estimate $\beta$ reliably based on the noisy data $y$?

To estimate $\beta$, we introduce a new estimator—we call it the Dantzig selector—which is a solution to the $\ell_1$-regularization problem:

$$
\min_{\beta \in \mathbb{R}^p} \| \tilde{\beta} \|_{\ell_1} \quad \text{subject to} \quad \| X^* r \|_{\ell_\infty} \leq \left( 1 + i^{-1} \right) \sqrt{2 \log p \cdot \sigma},
$$

where $r$ is the residual vector $y - X \tilde{\beta}$ and $i$ is a positive scalar. We show that if $X$ obeys a uniform uncertainty principle (with unit-normed columns) and if the true parameter vector $\beta$ is sufficiently sparse (which here roughly guarantees that the model is identifiable), then with very large probability,
George Bernard Dantzig

(Nov 8, 1914 - May 13, 2005)

An American mathematician

The father of linear programming

The inventor of “simplex method”
Central South University

The obverse of the Fields Medal

Terence Tao
Discussion of “the Dantzig selector”

Bradley Efron       Trevor Hastie       and Robert Tibshirani *

May 3, 2007

2 Dantzig selector and the lasso

The definition of the Dantzig selector (DS) in (1.7) can be re-expressed as

$$\min_{\beta} \|X^T(y - X\beta)\|_\infty \text{ subject to } \|\beta\|_{\ell_1} \leq s$$

(1)

This makes it look very similar to the lasso (Tibshirani 1996), or basis pursuit (Chen, Donoho & Saunders 1998):

$$\min_{\beta} \|y - X\beta\|_{\ell_2} \text{ subject to } \|\beta\|_{\ell_1} \leq s$$

(2)

With a bound on the \(\ell_1\) norm of \(\beta\), lasso minimizes the squared error while DS minimizes the maximum component of the gradient of the squared error function. If \(s\) is large so that the constraint has no effect, then these are the same. However for other values of \(s\), they are a little different; see Figure 1.
Geometry of DS

\[ e = y - \hat{y} \]

\[ r_2 = |\cos(\theta_2)| \]

\[ r_1 = |\cos(\theta_1)| \]

Green < Pink < Blue
## Discussion of “the Dantzig selector”

Bradley Efron, Trevor Hastie, and Robert Tibshirani

May 3, 2007

<table>
<thead>
<tr>
<th>Variable $j$</th>
<th>Lasso $X_j^T(y - X\hat{\beta})$</th>
<th>$\hat{\beta}_j$</th>
<th>Dantzig Selector $X_j^T(y - X\hat{\beta})$</th>
<th>$\hat{\beta}_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27.4134</td>
<td>0.0000</td>
<td>26.0046</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>-83.6413</td>
<td>-77.0062</td>
<td>-83.4945</td>
<td>-73.0993</td>
</tr>
<tr>
<td>3</td>
<td>83.6413</td>
<td>502.8695</td>
<td><strong>62.5323</strong></td>
<td><strong>543.7634</strong></td>
</tr>
<tr>
<td>4</td>
<td>83.6413</td>
<td>233.5998</td>
<td>83.4945</td>
<td>223.6250</td>
</tr>
<tr>
<td>5</td>
<td>-41.1153</td>
<td>0.0000</td>
<td>-43.5949</td>
<td>0.0000</td>
</tr>
<tr>
<td>6</td>
<td>-33.8190</td>
<td>0.0000</td>
<td>-37.0429</td>
<td>0.0000</td>
</tr>
<tr>
<td>7</td>
<td>-83.6413</td>
<td>-164.0632</td>
<td>-83.4945</td>
<td>-155.4648</td>
</tr>
<tr>
<td>8</td>
<td>51.2581</td>
<td>0.0000</td>
<td>50.5638</td>
<td>0.0000</td>
</tr>
<tr>
<td>9</td>
<td>83.6413</td>
<td>463.4805</td>
<td>83.4945</td>
<td>455.3289</td>
</tr>
<tr>
<td>10</td>
<td>83.6413</td>
<td>4.9767</td>
<td><strong>83.4945</strong></td>
<td><strong>0.0000</strong></td>
</tr>
<tr>
<td>11</td>
<td>76.1206</td>
<td>0.0000</td>
<td>75.6962</td>
<td>0.0000</td>
</tr>
<tr>
<td>12</td>
<td>83.6413</td>
<td>29.7423</td>
<td>83.4945</td>
<td>13.1410</td>
</tr>
</tbody>
</table>

| ... | ... | ... | ... | ... |

Table 1: Results for the Lasso and the Dantzig selector on the diabetes data with 64 variables (first 12 shown). The lasso and DS solutions have the same $l_1$ norm $||\hat{\beta}||_{l_1} = 1734.79$. 

中南大学化学化工学院中南现代化学研究中心
Discussion of “the Dantzig selector”

Bradley Efron    Trevor Hastie    and Robert Tibshirani

May 3, 2007

Figure 3: RMSE curves for lasso and DS for the simulation with $n = 15$, $p = 100$, and a sparse coefficient vector $\beta$ with 15 nonzero entries. The left panel uses a grid on $||\hat{\beta}||_1$, while the right on $||X^T(y - X\hat{\beta})||_\infty$. 

中南大学化学化工学院中南现代化研究中心
Conclusions from Efron

The optimality properties of the Dantzig selector established by the authors are impressive. We wonder if similar properties hold for the lasso, and hope that the authors can shed some light on this.

From our brief study, the inherent criterion in DS for including predictors in the model appears to be counter-intuitive, and its prediction accuracy seems to be similar to that of the lasso in some settings, and inferior in other settings. Hence we find little reason to recommend the Dantzig selector over the lasso.
Conclusions

- Variable selection is very important for building a high performance model
- Many open problems and open discussions
- Many methods are based on linear models
- Develop special methods for special data
Thanks so much!