LASSO-type regularization methods

Jinseog Kim

Department of Statistics & Information Science
Dongguk University

May. 21, 2008
Background

- In the regression problem, when the dimension of input is large (e.g. larger than the sample size), there are lots of problems in applying simple methods (e.g. least squares method).
- Most important problems in high dimensional problems are:
  - **Multicollinearity**
    - some input variables are highly correlated
    - For example, when the dimension of input is large than the sample size, the least squares estimator is not unique.
  - **Overfitting**: A model with too many input variables may be sub-optimal when the true model is sparse (a response variable depends only on a small number of input variables).
Possible remedies can be
- Variable selection: Best subset selection
- Regularization: Ridge, LASSO, SCAD, elastic net, etc.
- Dimension reduction techniques: Principal component regression, Partial least squares (not covered in this talk).
Variable Selection

- When there are $p$ input variables, we want to select the optimal model among all possible models.
- Let $M_k$ be a model whose residual mean squared error is minimum among all models having $k$ many input variables.
- Select the optimal model among $M_0, \ldots, M_p$.
- When $p$ is very large (say, larger than 100), this approach (all possible search) becomes computationally infeasible. ($2^{100} \approx 100,000$)
- There exists heuristics such as forward selection, backward elimination or stepwise, but such methods are still unstable.
Figure: best subset selection
Instability of variable selection

- Variable selection uses a hard decision rule (survive or die).
- So, it may result in inferior prediction error.
- ‘Unstable’ procedure means that small change of data results in large change of the estimator (eg. decision tree or neural net).
- Regularization methods are promising alternatives.
Regularization methods

- Ridge
- LASSO
- Extension of LASSO
1. Ridge regression

Ridge estimator $\hat{\beta}^{\text{ridge}}$ minimizes

$$\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

or equivalently

$$\sum_{i=1}^{n} l(y_i, \beta' x_i) + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (1)$$

where $\lambda > 0$ and $l(y, f) = (y - f)^2$ is a loss function.
1. Ridge regression (conti.)

- \( s \) (or \( \lambda \)): control or regularization parameter which controls the complexity of the model.
- If \( s = 0 \), the model includes only the intercept term while the model becomes the full model when \( s = \infty \).
- The ridge estimator was proposed by Hoerl and Kennard (1970) to solve the least squares problem when \( p > n \).
- Recall that the least squares estimator is given as

\[
\hat{\beta} = (X'X)^{-1}X'y.
\]

- When \( p > n \), \((X'X)^{-1}\) does not exist (ill-posed problem).
- By replacing \((X'X)^{-1}\) into \((X'X + \lambda I)^{-1}\),

\[
\hat{\beta} = (X'X + \lambda I)^{-1}X'y.
\]

- Note that the ridge estimator is the solution of (1).
Figure: Comparison of Selection and Ridge
**Table:** Comparison of Selection and Ridge

<table>
<thead>
<tr>
<th>Term</th>
<th>LS</th>
<th>Selection</th>
<th>Ridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.480</td>
<td>2.495</td>
<td>2.467</td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.680</td>
<td>0.740</td>
<td>0.389</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.305</td>
<td>0.367</td>
<td>0.238</td>
</tr>
<tr>
<td>$x_3$</td>
<td>-0.141</td>
<td>-0.029</td>
<td></td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.210</td>
<td></td>
<td>0.159</td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.305</td>
<td></td>
<td>0.217</td>
</tr>
<tr>
<td>$x_6$</td>
<td>-0.288</td>
<td></td>
<td>0.026</td>
</tr>
<tr>
<td>$x_7$</td>
<td>-0.021</td>
<td></td>
<td>0.042</td>
</tr>
<tr>
<td>$x_8$</td>
<td>0.267</td>
<td></td>
<td>0.123</td>
</tr>
<tr>
<td>Test Error</td>
<td>0.586</td>
<td>0.574</td>
<td>0.540</td>
</tr>
</tbody>
</table>
1. Ridge regression: Remarks

- The parameter $s$ or $\lambda$ is called the regularization parameter. The selection of this parameter is the same as model selection.
- We can extend the ridge estimator to logistic regression or NN easily by

$$
\beta_{\text{ridge}} = \arg \min_n \sum_{i=1}^{n} \log (1 + \exp(-y_i x_i' \beta)) + \lambda \sum_{k=1}^{p} \beta_k^2.
$$

- Regularized form of SVM is

$$
\min_\beta \sum_{i=1}^{n} (1 - y_i x_i' \beta)_+ + \lambda \sum_{k=1}^{p} \beta_k^2.
$$
2. LASSO

- A disadvantage of the ridge regression is that the interpretation is not easy since the final model includes all input variables.

- Can we do variable selection and shrinkage estimation simultaneously?

- The first of such methods is LASSO (Least Absolute Shrinkage and Selection Operator), firstly proposed by Tibshirani (1996).
2. LASSO

- LASSO estimates $\beta$ by minimizing

$$\sum_{i=1}^{n} l(y_i, x_i^\prime \beta) + \lambda \sum_{j=1}^{p} |\beta_j|.$$ 

- The only difference to ridge is the penalty function.
- We can say that the $l_1$ penalty is used in LASSO while the $l_2$ penalty is used in ridge.
- This seemingly tiny difference makes qualitative gaps practically as well as theoretically.
- One very interesting property of LASSO is that the predictive model is sparse (i.e. some coefficients are exactly 0).
Figure: Solution paths of ridge and LASSO for Prostate Cancer data
Table: Illustration: Why can LASSO get a sparse solution?
# LASSO v.s the other methods

<table>
<thead>
<tr>
<th>Term</th>
<th>LS</th>
<th>Selection</th>
<th>Ridge</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.480</td>
<td>2.495</td>
<td>2.467</td>
<td>2.477</td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.680</td>
<td>0.740</td>
<td>0.389</td>
<td>0.545</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.305</td>
<td>0.367</td>
<td>0.238</td>
<td>0.237</td>
</tr>
<tr>
<td>$x_3$</td>
<td>-0.141</td>
<td>-0.029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.210</td>
<td>0.159</td>
<td>0.098</td>
<td></td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.305</td>
<td>0.217</td>
<td>0.165</td>
<td></td>
</tr>
<tr>
<td>$x_6$</td>
<td>-0.288</td>
<td>0.026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_7$</td>
<td>-0.021</td>
<td>0.042</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_8$</td>
<td>0.267</td>
<td>0.123</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>Test Error</td>
<td>0.586</td>
<td>0.574</td>
<td>0.540</td>
<td>0.491</td>
</tr>
</tbody>
</table>

LASSO solution can achieve both selection and shrinkage!
Computation of LASSO

- We can see that a key property of sparse penalty function is that it is nondifferentiable around 0.
- That is, for sparse learning, we need to optimize a nondifferentiable objective function.
- Hence, standard numerical optimization methods can not be applied directly.
- Since LASSO was proposed firstly, optimization problem has been one of the hottest issues in statistics and machine learning society.
- Three algorithms: one based on the QP, the second based on angle, and the last one based on gradient descent.
Computation of LASSO (conti.)

- LASSO problem:

\[
\min \sum_{i=1}^{n} l(y_i, x_i' \beta) \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s \text{ for some } s > 0.
\]

- When \( l(y, a) = (y - a)^2 \), this is a quadratic programming problem with linear constraints, and so we can apply any QP algorithm, which was done by Tibshirani (1996).
- Osborne (2000a, 2000b) and Rosset and Zhu (2007) developed more efficient path finding algorithms, but it can be only applicable to piecewise quadratic losses.
- Gradient LASSO algorithm (Kim et al. 2005): an optimization algorithm for general convex losses.
The gradient LASSO algorithm

1. Initialize: $w = 0$ and $m = 0$
2. Do until converge:
   1. Coordinate Gradient Descent (Addition step)
   2. Deletion algorithm
   3. $m = m + 1$
3. Return $w$
CGD algorithm

CGD algorithm is to update $\beta$ sequentially as follows.

1. Compute the gradient vector $\nabla(\beta) = \partial C(\beta)/\partial \beta$.
2. Find the coordinate vector $u \in E$ which minimizes $\langle \nabla(\beta), u \rangle$.
3. Find the convex combination which minimizes

$$C\left((1 - \alpha)\beta + \alpha u\right).$$

4. Update $\beta \leftarrow (1 - \alpha)\beta + \alpha u$. 
CGD algorithm (1)

1. Let $\beta^{(1)}$ be a current solution. 2. Find the coordinate vector (vertex). 3. Take convex combination.
CGD algorithm (2)

1. Move to optimal convex combination. 2. Find the coordinate vector and take convex combination. 3. Move to optimal convex combination.
Gradient LASSO algorithm

Typical path

CGD

Deletion(1)

Deletion(2)
Extension of the LASSO

- Group LASSO (Yuan and Li, 2006), Blockwise Sparse Regression (Kim et al. 2006): LASSO method for selecting groups of variables.
- Elastic net (Zou and Hastie, 2005): a shrinkage method for yielding less sparse solution than LASSO.
- SCAD (Fan and Li, 2001): a new shrinkage method which has an oracle property.
Group LASSO or BSR

- Consider a grouping structure on the covariates.
- Categorical variable with \( k \) levels.
  - An usual technique is to use \( k - 1 \) dummies.
  - LASSO may select some of dummies and drop the others.
- \( k \)-th polynomial fit from a numeric covariate.
  - An usual technique is to use \( k \) terms from the power transformations.
  - LASSO may select some of terms and drop the others.
- Individual sparsity does NOT ensure blockwise sparsity.
- Group LASSO was developed to ensure the blockwise sparsity. That is, all coefficients in the same blocks are either all nonzero or all zero.
- Group LASSO is firstly proposed by Yuan and Li (2006) for linear regressions.
- Kim et al. (2006) proposed a similar method (Blockwise Sparse Regression), which deals with general convex loss functions.
Group LASSO or BSR: Formulation

- A block structure on coefficients: \( \mathbf{\beta} = (\mathbf{\beta}'_1, \cdots, \mathbf{\beta}'_d)' \)
- Motivation: Ridge penalty can be re-expressed as

\[
J(\mathbf{\beta}) = \sum_{j=1}^{d} \|\mathbf{\beta}_{(j)}\|_2^2
\]

- The grouped LASSO finds a solution from

\[
\text{minimizing } \sum_{i=1}^{n} L(y_i, \mathbf{\beta}'x_i) \text{ subject to } \sum_{j=1}^{d} \|\mathbf{\beta}_{(j)}\|_2 \leq M
\]

- Property: Compromising Ridge and Lasso.
  - Blockwise Sparsity.
  - Ridge type Shrinkage within a block.
Group LASSO or BSR: Illustration

**Figure:** LASSO, group LASSO and Ridge† (Yuan and Lin, 2006)

\[
\sqrt{\beta_{11}^2 + \beta_{12}^2 + \beta_2^2} \leq M
\]
Group LASSO or BSR: Computation

- Modified LASSO algorithm (Yuan and Li, 2006)
- Gradient projection algorithm (Kim et al. 2006)
- Zho, Rocha and Yu (2006) extended the group lasso when there is hierarchical structure within variables (e.g. wavelet approximation).
Elastic Net: Motivation

- Zou and Hastie (2005, JRSS-B)

- Sparse methods are good only when the true model is sparse.

- When there are high correlation between covariates, the average of the covariates would be better than the selection of a covariate.

- An example is a factor model.
  - True model: $Y = F + \epsilon$ where $F \sim N(0, 1) \perp \epsilon_i \sim N(0, \sigma^2)$.
  - Data: $(Y, X_1, X_2)$ where $X_j = F + \xi_i$ and $\xi_j \sim N(0, 1) \perp F$.
  - Then

$$\arg\min_{\beta_1, \beta_2} E(Y - \beta_1 X_1 - \beta_2 X_2)^2 = (1/2, 1/2).$$

- So we need a method of simultaneously automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. The elastic net is such a method.
Elastic Net: Illustration

Figure: Penalties for LASSO, ridge, and elastic net
Elastic Net: Computation

- The main idea of the elastic net is to combine the ridge and Lasso.
- *Naive elastic net* minimizes the following objective function:

\[ \sum_{i=1}^{n} (y_i - \beta' x_i)^2 + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_2^2. \]

- A problem: naive elastic net estimator needs to be doubly regularized.
Elastic Net: Computation (conti.)

- Given $\lambda_2$, let $X^* = (1 - \lambda_2)^{-1/2} \left( \frac{X}{\sqrt{\lambda_2 I}} \right)$, and $y^* = \begin{pmatrix} y \\ 0 \end{pmatrix}$,

  $${\hat{\beta}}^* = \text{arg min} \ |y^* - X^* \beta^*|^2 + \frac{\lambda_1}{\sqrt{1 + \lambda_2}} \|\beta^*\|_1.$$ 

  $${\Rightarrow {\hat{\beta}} = (1 + \lambda_2)^{-1/2} {\hat{\beta}}^*}$$

- Above is stagewise approach which often results in poor performance. To overcome such deficiency, Zou and Hastie (2005) proposed an ad-hoc modification:

  $${{\hat{\beta}}(\text{elastic net}) = (1 + \lambda_2) {\hat{\beta}}(\text{naive elastic net}).}$$
Elastic Net: Illustration (2)

**Figure:** Solutions for LASSO, ridge, and elastic net under orthogonal design
Elastic Net: Illustration (3)

Figure: solution path of LASSO and elastic net
SCAD (Smoothly Clipped Absolute Deviation): Idea

- Fan and Li (2001) generalized the penalized approaches:

\[
\hat{\beta} = \arg\min_\beta C_n(\beta) + \sum_{j=1}^{p} J_\lambda(\|\beta_j\|),
\]

where \( J \) is a penalty function and
\[
C_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i' \beta)^2.
\]

- Three desirable properties for penalty functions:
  - Unbiasedness: of the estimator
  - Sparsity: Some coefficients are exactly zero to reduce model complexity.
  - Continuity: The estimator is continuous in data to avoid instability.
SCAD: penalty functions

- Hard thresholding (Fan, 1997): $\lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$
- Bridge: $J_\lambda(\theta) = \lambda|\theta|^q, q > 0$
- Lasso: $J_\lambda(\theta) = \lambda|\theta|$
- Ridge: $J_\lambda(\theta) = \lambda|\theta|^2$
- SCAD:

\[
J_\lambda(\theta) = \begin{cases} 
\lambda|\theta|, & |\theta| \leq \lambda, \\
-(\theta^2 - 2a\lambda|\theta| + \lambda^2)/[2(a - 1)], & \lambda \leq |\theta| \leq a\lambda \\
(a + 1)\lambda^2/2, & |\theta| \geq a\lambda.
\end{cases}
\]
SCAD: Illustration

Figure: LASSO, ridge, and SCAD penalties
## SCAD: Properties

### Table: Penalties and their properties

<table>
<thead>
<tr>
<th></th>
<th>Bridge((q &lt; 1))</th>
<th>LASSO</th>
<th>Ridge</th>
<th>SCAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>sparsity</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>O</td>
</tr>
<tr>
<td>unbiasedness</td>
<td>O</td>
<td>X</td>
<td>X</td>
<td>O</td>
</tr>
<tr>
<td>continuity</td>
<td>X</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
**SCAD: Oracle property**

- Let $Y_n = X_n\beta^* + \epsilon_n$ where $\epsilon_n = (\epsilon_1, \ldots, \epsilon_n)$ is a vector of i.i.d. random variables with mean 0, $Y_n = (y_1, \ldots, y_n)'$ and $X_n = (X_n^1, \ldots, X_n^p)$ with $X_n^j = (x_{jn1}, \ldots, x_{jnn})'$.

- Suppose $\beta^* = (\beta^{(1)*}, \beta^{(2)*})$ where $\beta^{(1)*}$ is the $q$-dimensional vector of non-zero coefficients and $\beta^{(2)*}$ is the last $p - q$ zero regression coefficients. Also, let $X_n = (X_n^{(1)}, X_n^{(2)})$.

- Let $\hat{\beta}^o = (\hat{\beta}^{(1)o}, 0)$ be the oracle estimator where $0$ is the $p - q$ dimensional 0-vector and

$$\hat{\beta}^{(1)o} = \arg\min_{\beta^{(1)}} \|Y_n - X_n^{(1)}\beta^{(1)}\|_2^2.$$  

- That is, the oracle estimator is a least square estimator with knowing zero coefficients in advance.
SCAD: Oracle property (conti.)

- We say an estimator $\hat{\beta}$ has the oracle property if it is asymptotically equivalent to the oracle estimator.

- When $p$ is fixed, there are the two regularized estimators which have oracle properties
  - SCAD (Fan and Li, 2001)
  - Adaptive Lasso (Zou, 2006)

- When $p \to \infty$ but $p/n \to 0$, the SCAD estimator is only one known to have the oracle property (Fan and Peng, 2004, Annals).

- Recently, Kim et al. (2007) proved that the SCAD estimator has the oracle property even when $p_n = O(\exp(an))$. 
## Simulation: linear regression and logistic regression

<table>
<thead>
<tr>
<th>method</th>
<th>n=40, $\sigma = 3$</th>
<th>n=60, $\sigma = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRME</td>
<td>Correct</td>
</tr>
<tr>
<td>SCAD1</td>
<td>72.90</td>
<td>4.37</td>
</tr>
<tr>
<td>SCAD2</td>
<td>69.03</td>
<td>4.42</td>
</tr>
<tr>
<td>LASSO</td>
<td>73.82</td>
<td>3.56</td>
</tr>
<tr>
<td>Ridge</td>
<td>83.28</td>
<td>0</td>
</tr>
<tr>
<td>Best subset</td>
<td>68.26</td>
<td>4.50</td>
</tr>
<tr>
<td>Oracle</td>
<td>33.31</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MRME</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCAD (a= 307)</td>
<td>26.48</td>
<td>4.98</td>
<td>0.04</td>
</tr>
<tr>
<td>LASSO</td>
<td>53.14</td>
<td>3.76</td>
<td>0</td>
</tr>
<tr>
<td>Hard</td>
<td>59.06</td>
<td>4.27</td>
<td>0</td>
</tr>
<tr>
<td>Best subset</td>
<td>31.63</td>
<td>4.84</td>
<td>0.01</td>
</tr>
<tr>
<td>Oracle</td>
<td>25.71</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
SCAD: Computation

- The objective function of the SCAD is non-convex, and hence we need a special optimization algorithm.

- Fan and Li (2001) proposed a crude linear approximation of the objective function, which may not converge to (local) minimum or fails to converge when the dimension of input is large.

- Recently, Kim et al. (2007) developed an algorithm, which always converges to a (local) minimum by combing the CCCP (Convex-ConCave Procedure) and Lasso algorithm.
SCAD: Computation (conti.)

- Convex-Concave Procedure (CCCP, Yille and Rangarajan 2003)
- Suppose we are to minimize a non-convex function $C'(\beta)$.
- Suppose $C'(\beta)$ is a sum of convex and concave functions $C_{vex}(\beta)$ and $C_{cav}(\beta)$ such as

$$C(\beta) = C_{vex}(\beta) + C_{cav}(\beta).$$

- For a given current solution $\beta^c$, the tight convex upper bound is defined by $Q(\beta) = C_{vex}(\beta) + \nabla C_{cav}(\beta^c)' \beta$ where

$$\nabla C_{cav}(\beta) = \partial C_{cav}(\beta)/\partial \beta.$$

- We update the solution by the minimizer of $Q(\beta)$ which is convex.

- Repeat this until convergence. It always converges to a local minimum.
SCAD: Computation (conti.)

Decomposition of the SCAD penalty

- Let $\tilde{J}_\lambda(|\beta_j|) = J_\lambda(|\beta_j|) - |\beta_j|$.
- Then, $\tilde{J}_\lambda(|\beta_j|)$ is a smooth concave function.
SCAD: Computation (conti.)

The CCCP-LASSO algorithm

- The algorithm is a hybrid of the CCCP and Lasso algorithm.
- Write

\[ C_n(\beta) = \|Y_n - X_n\beta\|^2_2/n + \lambda \sum_{j=1}^{p} \sum_{\tilde{J}_j} \tilde{J}_\lambda(|\beta_j|) + \lambda \sum_{j=1}^{p} |\beta_j|. \]

- Then, for given current solution \( \beta^c \), the convex upper bound is given as

\[ Q(\beta) = \|Y_n - X_n\beta\|^2_2/n + \lambda \sum_{j=1}^{p} \nabla \tilde{J}_\lambda(|\beta^c_j|)' \beta + \lambda \sum_{j=1}^{p} |\beta_j| \]

- \( Q(\beta) \) is easily minimized by use of the Lasso algorithm developed by Rosset and Zhu (2007, Annals).
SCAD: Remark

- The CCCP-LASSO algorithm can be applied to piecewise quadratic loss functions.
- However, it is not directly applicable to general convex losses such as logistic regressions.
- Also, the oracle property for general convex loss remains to be proved.
Summary

- Regularization is an important method to increase predictive accuracy and interpretation of the model, in particular when $n < p$. In recent years, lots of regularization methods have been proposed.

- It leaves much room for improvement both computationally and theoretically.
References


References


References


