The Elements of Statistical Learning

Ch3: Liner Methods for Regression - Shrinkage Methods

Philip Lin

Data Science in Hsinchu

2015.08.05

Framework

Today, we talk about one famous shrinkage methods: Bridge family

$$\arg\min_{\beta}\Big\{\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j|^q\Big\} \quad ,q\geq 0$$

we discuss some cases

- $q=0$: Best subset selection (Garside, 1965)
- $q=2$: Ridge regression (Hoerl and Kennard, 1970)
- $q=1$: Lasso (Tibshirani, 1997)
- $1<q<2$: Elastic Net (Hui Zou, 2006)

Keywords:

Shooting algorithm, Oracle property, SCAD, Adaptive weight, Group effect

Best subset selection

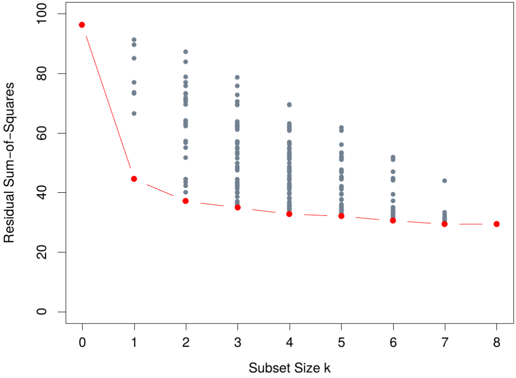
$$\arg\min_{\beta}\Big\{\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j|^0\Big\}$$

- Here, we define $|\beta|^0 = I\{\beta \neq 0\}$
- By Lagrangian, we have equivalent objective

$$\arg\min_{\beta}\Big\{\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2\Big\} \quad \text{s.t.} \quad \sum_{j=1}^{p}I\{\beta_j\neq 0\}\leq t, \quad t\geq 0$$

- Denote $k=\{0,1,\dots,p\}$ as number of covariates included in the model. Best subset selection finds **for each** k, the subset of size k, that gives smallest residual sum of squares.
- An efficient algorithm – *leaps and bounds* procedure makes this feasible for p as large as 30 or 40.

e.g.: $p = 8$



- ▷ The best curve is necessary decreasing, so cannot be used to select subset size k. There are many choices of choosing proper k, such as AIC, BIC, CV etc.

The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

Ridge

Consider the following loss function

$$\begin{aligned}\phi =&(y - XB)^T(y - XB)\\ =&(y - X\hat{\beta})^T(y - X\hat{\beta}) + (B - \hat{\beta})^T X^T X(B - \hat{\beta})\\ =&\phi_{\min} + \phi(B)\end{aligned}$$

where $\hat{\beta}$ is the minimizer of ϕ (i.e., LSE) and here B is any estimator of β

Remarks:

- If we aim to minimize ϕ, LSE is the Best Linear Unbiased Estimator (BLUE) which is the famous *Gauss Markov Theory*.
- LSE is unbiased and has minimum variance among all linear unbiased estimators.
- In other words, there's something else that has smaller variance than LSE outside of the class of linear unbiased estimators.

Let's consider the squared distance from $\hat{\beta}$ to β, then the expected distance is

$$\begin{aligned} E(\hat{\beta}-\beta)^T(\hat{\beta}-\beta) =& \text{tr}\left\{E(\hat{\beta}-\beta)^T(\hat{\beta}-\beta)\right\} \\ =& \text{tr}\left\{E\hat{\beta}^T\hat{\beta}-\beta^T\beta\right\} \\ =& E\left\{\text{tr}(\hat{\beta}^T\hat{\beta})\right\}-\beta^T\beta \\ =& \sigma^2\text{tr}\{(X^TX)^{-1}\} \\ =& \sigma^2\sum_{j=1}^{p}\frac{1}{a_i} \end{aligned}$$

or equivalently

$$E[\hat{\beta}^T\hat{\beta}] = \beta^T\beta+\sigma^2\text{tr}\{(X^TX)^{-1}\}$$

Note

- the eigenvalues of X^TX are denoted by

$$a_{\max}=a_1\geq a_2\geq\cdots\geq a_p=a_{\min}>0$$

- If there's one or more small eigenvalues, the squared distance from $\hat{\beta}$ to β will tend to be large. (Recall PCA and multiple collinearity)

The main idea

- Think of ϕ as the surface of hyperellipsoids centered at $\hat{\beta}$, LSE of β.

  The average distance from $\hat{\beta}$ to β can be writen as

$$E(\hat{\beta}-\beta)^T(\hat{\beta}-\beta) = E[\hat{\beta}^T\hat{\beta}] - \beta^T\beta = \sigma^2 \text{tr}\{(X^TX)^{-1}\}$$

  The average distance tend to be large if there is a small eigenvalue of X^TX

- We should move away from LSE, but how to decide the direction ?

$\Rightarrow$ From the view of

$$\begin{aligned} E[\hat{\beta}^T\hat{\beta}] =& \beta^T\beta + \sigma^2 \text{tr}\{(X^TX)^{-1}\} \\ E[B^TB] =& \beta^T\beta + \text{ something esle} \end{aligned}$$

we expect that

$$E[B^TB] \le E[\hat{\beta}^T\hat{\beta}]$$

that is, the movement should be in a direction which will **shorten the length** of the regression vector.

The question becomes

$$\min_{\beta} (y - X\beta)^T (y - X\beta) \quad \text{subject to } \beta^T\beta \leq t \quad , t \geq 0$$

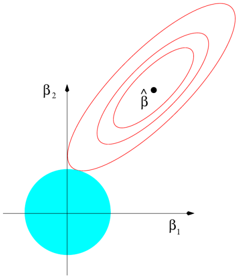
or equivalently

$$\min_{\beta} \left\{ (y - X\beta)^T (y - X\beta) + \lambda\beta^T\beta \right\} \quad , \lambda \geq 0$$

where λ is one to one mapped to t

$\Rightarrow$ The **Ridge** estimate

$$\hat{\beta}^* = (X^T X + \lambda I)^{-1} X^T y$$



The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

The construction of L_2 penalized least square problem is based on reducing squared length between estimator $\hat{\beta}$ and the truth β. It is worthwhile to check the performance of Ridge.

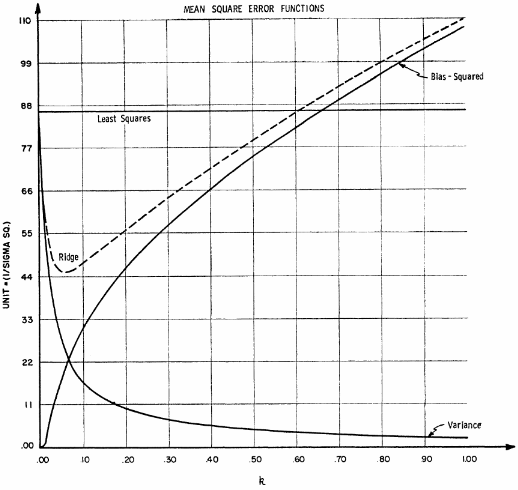
Define $\hat{\beta}^* = (X^T X + \lambda I)^{-1}(X^T X)\hat{\beta} = Z\hat{\beta}$ such that $E(\hat{\beta}^*) = Z\beta$

$$\begin{aligned}
E\Big[(\hat{\beta}^* - \beta)^T(\hat{\beta}^* - \beta)\Big] =& E\Big[(Z\hat{\beta} - Z\beta + Z\beta - \beta)^T(Z\hat{\beta} - Z\beta + Z\beta - \beta)\Big] \\
=& E\Big[(\hat{\beta} - \beta)^T Z^T Z(\hat{\beta} - \beta)\Big] + (Z\beta - \beta)^T(Z\beta - \beta) \\
=& \sigma^2 \mathrm{tr}\Big[(X^T X)^{-1} Z^T Z\Big] + \beta^T(Z - I)^T(Z - I)\beta \\
=& \sigma^2 \sum_{j=1}^{p} \frac{a_j}{(a_j + \lambda)^2} + \lambda^2 \beta^T (X^T X + \lambda I)^{-2}\beta \\
=& \gamma_1(\lambda) + \gamma_2(\lambda) \\
=& \text{Variance}^2 + \text{Bias}^2
\end{aligned}$$

it can be shown that

- $\gamma_1(\lambda)$ is a continuous, monotonically decreasing function of λ.
- $\gamma_2(\lambda)$ is a continuous, monotonically increasing function of λ.
- There always exist a $\lambda > 0$ such that

$$E\Big[(\hat{\beta}^* - \beta)^T(\hat{\beta}^* - \beta)\Big] < E\Big[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)\Big]$$



The figure is cited from "Hoerl, A.E. & Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), pp.55 - 67."

Let us see Gauss Markov again

$$\begin{aligned}\phi =&(y - XB)^T(y - XB)\\ =&(y - X\hat{\beta})^T(y - X\hat{\beta}) + (B - \hat{\beta})^T X^T X(B - \hat{\beta})\\ =&\phi_{\min} + \phi(B)\end{aligned}$$

A completely equivalent statement of ridge optimization problem can be stated as

Minimize $B^T B$

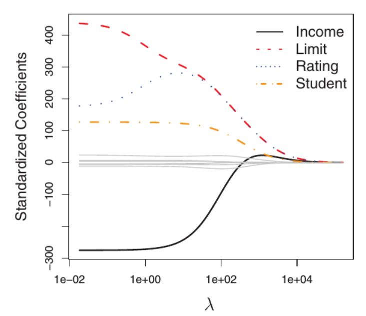
$$\text{subject to } (B - \hat{\beta})^T X^T X(B - \hat{\beta}) \leq t$$

By Lagrangian, let $F = B^T B + (1/\lambda)[(B - \hat{\beta})^T X^T X(B - \hat{\beta})]$

$$\begin{aligned}\frac{\partial F}{\partial B} =&2B + (1/\lambda)[2(X^T X)B - 2(X^T X)\hat{\beta}] = 0\\ B =&\hat{\beta}^* = (X^T X + \lambda I)^{-1} X^T y\end{aligned}$$

Remarks

- We move away from $\hat{\beta}$, but we hope not to be too far away from it.
- In this view, we expect that the performance of $(y - X\hat{\beta}^*)^T(y - X\hat{\beta}^*)$ will not be too bad.



Remarks

- ▷ The model complexity was embeded into λ (Compare to L_0 penalty). Thus, model selection problem becomes how to select a good λ.
- ▷ take a look at the Ridge **solution path**
  - All estimates are nonzero.
  - Since the shrinkage effect, estimate of irrelavent covariantes are closed to zero (gray lines). Why can you say that ?
  - Is it possible to automatically exclude the gray lines from our model ?

The figure is cited from "James, G. et al., 2013. An Introduction to Statistical Learning with Applications in R"

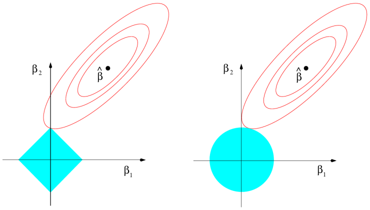
Lasso

To obtain **sparity** in coefficient estimate, Tibshirani (1996) proposed *Least Absolute Shrinkage and Selection Operator*. The question becomes

$$\arg\min_{\beta} \ (y - X\beta)^T(y - X\beta) \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \le t \quad , t \ge 0$$

or equivalently

$$\arg\min_{\beta} \left\{ (y - X\beta)^T(y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad , \lambda \ge 0$$



The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

Let's solve the L_1 problem coordinatewisely,

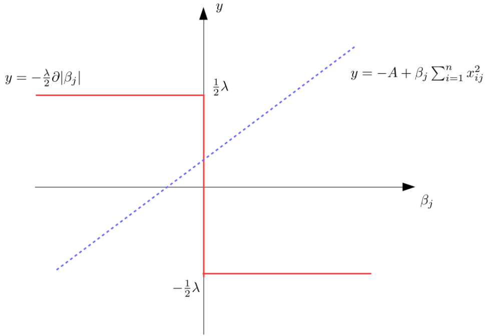
$$
\begin{aligned}
f =& \sum_{i=1}^{n}(y_i - \sum_{k \neq j} x_{ik}\tilde{\beta}_k - x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \\
\frac{\partial f}{\partial \beta_j} =& -2\sum_i (y_i - \sum_{k \neq j} x_{ik}\tilde{\beta}_k - x_{ij}\beta_j)x_{ij} + \lambda \partial |\beta_j| \\
=& -2\sum_i (y_i - \sum_{k \neq j} x_{ik}\tilde{\beta}_k)x_{ij} + 2\sum_i x_{ij}^2 \beta_j + \lambda \partial |\beta_j| \\
=& -2A + 2\beta_j \sum_i x_{ij}^2 + \lambda \partial |\beta_j|
\end{aligned}
$$

where

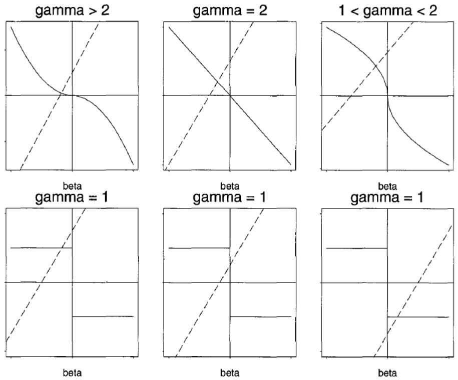
$$
\partial |\beta_j| = \begin{cases} \mathrm{sgn}(\beta_j) & \text{if} \quad \beta_j \neq 0 \\ \in (-1, 1) & \text{if} \quad \beta_j = 0 \end{cases}
$$

Observe the **KKT** condition:

$$-A + \beta_j \sum_i x_{ij}^2 = -\frac{\lambda}{2}\partial|\beta_j|$$



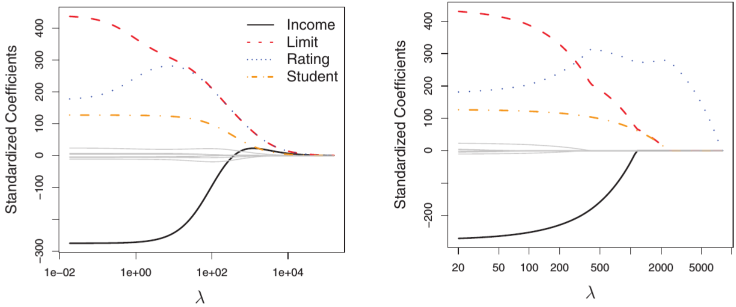
- The solution $\hat{\beta}_j$ is intersection of blue and red line.
- It can be shown that L_1 optimization problem can be solved coordinatewisely. (Shooting algorithm. Wenjiang J. Fu, 1998; Coordinate Descent. Friedman, 2006)



- In Bridge family, for $1 \leq q \leq 2$, only $q = 1$ has sparsity.

The figure is cited from "Fu, W.J., 1998. Penalized Regressions : The Bridge Versus the Lasso. Journal of Computational and Graphical Statistics, 7(3), pp.397 - 416."

Solution path



- Sparsity can be easily obtained by introducing non-differentiability at $\beta = 0$.
- Due to the sparsity, model selection can be regarded as a **large enough** λ such that some of irrelavent coefficient were estimated as zero.
- However, as λ increase, bias of nonzero estimates also increase.

The figure is cited from "James, G. et al., 2013. An Introduction to Statistical Learning with Applications in R"

Bias of Lasso problem

- For the nonzero set of coefficient estimate, $\mathcal{A}$, by KKT condition we have

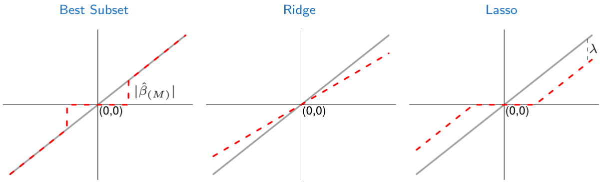
$$
\begin{aligned}
0 =& -2X_{\mathcal{A}}^T y + 2X_{\mathcal{A}}^T X_{\mathcal{A}} \hat{\beta}_{\mathcal{A}}^{\text{Lasso}} + \lambda \text{sgn}(\hat{\beta}_{\mathcal{A}}^{\text{Lasso}}) \\
\hat{\beta}_{\mathcal{A}}^{\text{Lasso}} =& (X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} \left(X_{\mathcal{A}}^T y - \frac{1}{2} \lambda \text{sgn}(\hat{\beta}_{\mathcal{A}}^{\text{Lasso}}) \right)
\end{aligned}
$$

  cf. ridge estimator

$$\hat{\beta}^* = (X^T X + \lambda I)^{-1} (X^T y)$$

  It is obvious to see that both of Ridge and Lasso are BIASED estimator.
- For simplicity, consider a special case: $X^T X = I$

$$\hat{\beta}^* = \frac{1}{1+\lambda} \hat{\beta}, \qquad \hat{\beta}^{\text{Lasso}} = \text{sgn}(\hat{\beta}) \left(|\hat{\beta}| - \frac{1}{2}\lambda \right)_+$$



The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

Remarks

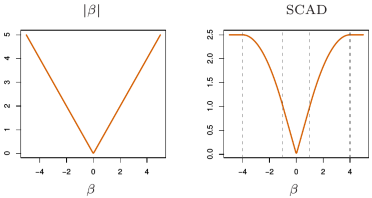
- We need to choose λ **large enough** so that irrelavent covariates obtain zero estimate

  We also need to choose λ **not too large** to avoid large bias in non zero estimate.
- You can use LSE to restimate the covariates selected by Lasso, however, it deviate the sprits of bias variance trade off.
- How to choose λ ? or What is a good model selection criterion ?
- Fan Jianqing (2001) claimed that a good penalty function should result in an estimator with three properties: *Unbiasedness*, *Sparsity* and *Continuity*.

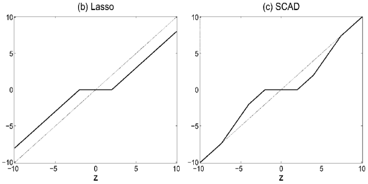
Fan Jianqing (2001)

- To attain these three properties, he proposed SCAD penalty, $J(\beta)$, instead of L_1

$$J'(\beta) = \lambda \text{sgn}(\beta)\left[I(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)\right]$$



- In orthogonal case



The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608.", "Fan, J. & Li, R., 2001. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association, 96(456), pp.1348-1360."

Oracle property

1. In the adjustment of penalty, bias of large coefficient vanished asymptotically. In otherwords, all nonzero coefficient estimates are as efficient as MLE when n is large enough.
2. Suppose λ is related to n and we denote it as λ_n. (actually, it does.) Fan Jianqing (2001) showed that asymptotically, there exist a range of λ_n such that all irrelavent coefficient will be estimated as zero.

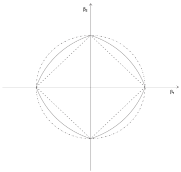
1. and 2. are summarized as **Oracle property**, which has an important effect on this decade.

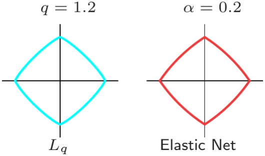
Elastic Net

$$\arg\min_{\beta}\left\{\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j|^q\right\}\quad, q\geq 0$$

- Till know, we discussed $q=2$ and $q=1$, what do we have for $1\leq q\leq 2$?
- Not interested. Why ? (do we have sparsity in $1<q\leq 2$?)
- Zou Hui (2005) approximate $1\leq q\leq 2$ by a combination of L_1 and L_2

$$\arg\min_{\beta}\left\{\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\left(\alpha\sum_{j=1}^{p}|\beta_j|+(1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)\right\}$$





- interestingly, we do have sparsity in Elastic Net

The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, Available at: http://www.springerlink.com/index/10.1007/b94608."

To be continued...