From ridge and lasso to elastic net

Hsieh Yuan-Hao

2024/11/08

1011 Recap(Effective degrees of freedom)

Degrees of Freedom in Ridge Regression:

$$\mathrm{df}(\lambda) = \mathrm{tr}\left[\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\right]$$

- This is monotone decreasing function of λ.($\lambda \uparrow \mathrm{df}(\lambda) \downarrow$)
- Usually in a linear-regression fit with p variables, the degrees-of-freedom of the fit is p.
- Although all p coefficients in a ridge fit will be non-zero,they are fit in restricted way controlled by λ
- $\mathrm{df}(\lambda) = p$ when $\lambda = 0$ (no regularization) and $\mathrm{df}(\lambda) \to 0$ as $\lambda \to \infty$

1011 Recap (Shrinkage Factor s in Lasso Regression)

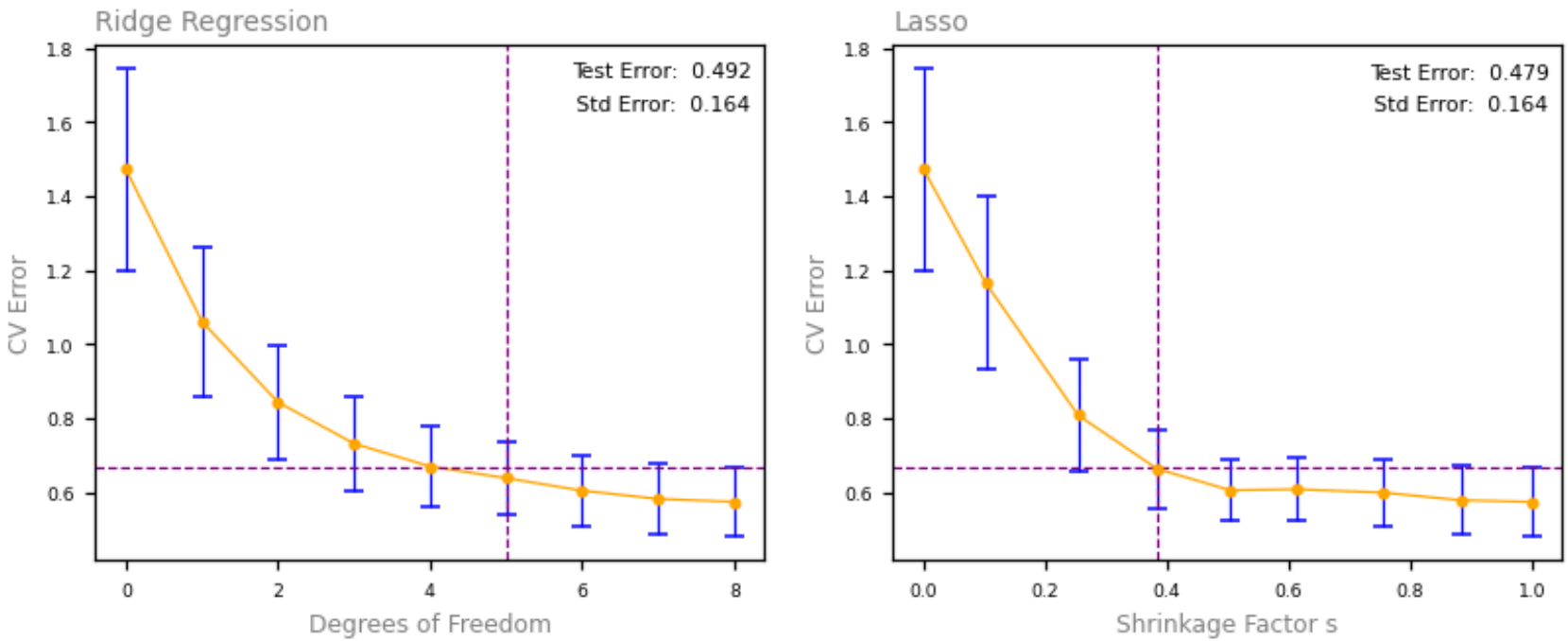
Shrinkage Factor s in Lasso Regression:

$$s = \frac{t}{\sum_{j=1}^{p} |\hat{\beta}_j|}$$

Where:

- t is the constraint on the sum of the absolute values of the coefficients $\hat{\beta}_j = \hat{\beta}_j^{ls}$; $\sum_{j=1}^{p} |\hat{\beta}_j|$ is the sum of the absolute values of the least-squares estimates of the coefficients.
- when t is chosen larger than $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$ (s$\geq$1), then the lasso estimates are the $\hat{\beta}_j$'s. (no sharinkage)
- If $t = t_0/2$ then the least squares coefficients are shrunk by about 50%, which means $s = 0.5$.
- $s = 1.0$ these are least squares estimates; these estimates decrease to 0 as $s \to 0$

1011 Recap(Ridge and lasso CV Error plot)



(a) Ridge regression

(b) Lasso regression

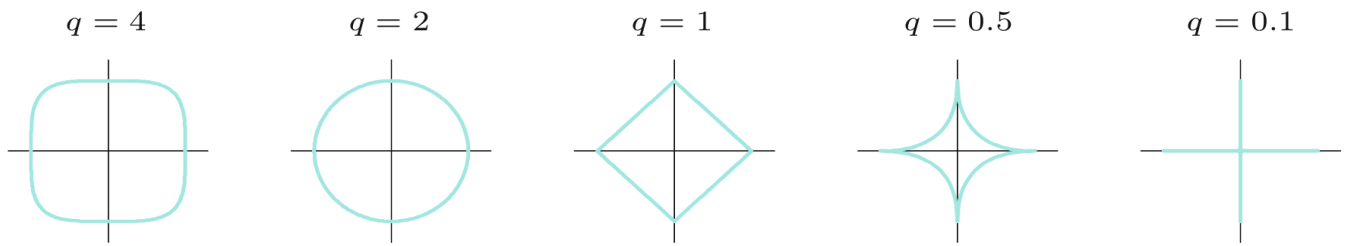
Figure: Compare between ridge and lasso

Motivation to introduced the *elastic-net* (1/2)

We can generalize ridge regression and lasso regression, Consider the criterion:

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

- The value $q = 0$ corresponds to variable subset selection.
- $q = 1$ corresponds to the lasso,while $q = 2$ to ridge regression.
- Using other values of q besides 0,1 or 2.

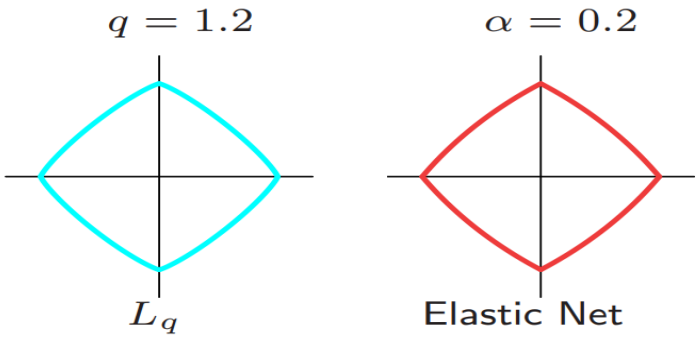


Motivation to introduced the *elastic-net* (2/2)

- Values of $q \in (1, 2)$suggest a compromise between the lasso and ridge regresison.
- With $q > 1, |\beta_j|^p$ is differentiable at 0, so this doesn't share the ability of lasso($q = 1$),setting coefficient exactly to zero.
- For this reason, Zou and Hastie(2005) introduced the *elastic net* penalty:

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1 - \alpha)|\beta_j|)$$

a different compromise between ridge and lasso.



- The elastic-net has sharp (non-differentiable) corners.
- The $q = 1.2$ penalty does not.

Introduction (Three scenarios)

Lasso has some limitations in the following scenarios:

(a) p>n Case: When the number of predictors p is much larger than the number of observations n the lasso can select at most n predictors.

(b) Grouped Variables:If there is a group of predictors with high pairwise correlations, the lasso tends to select only one predictor from the group, without preference for any specific one. This behavior ignores potentially useful grouping information.

(c) High Correlation: In cases where predictors are highly correlated, ridge regression has been observed to outperform the lasso in prediction accuracy. Thus, the lasso may not be the best choice for prediction in these situations due to weaker performance.

Naive elastic net

Assume that the response is centred and the predictors are standardized

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \textit{for } j = 1, 2, ..., p \quad (1)$$

For any fixed non-negative λ_1 and λ_2, we define the navie elastic net criterion

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1 \quad (2)$$

where

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2 \quad |\beta|_1 = \sum_{j=1}^{p} |\beta_j|$$

The navie elastic net estimator $\hat{\beta}$ is the minimizer of equation(3):

$$\hat{\beta} = \arg\min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\} \quad (3)$$

Naive elastic net

This procedure can be viewed as penalized least squares method.Let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$then solving $\hat{\beta}$ in equation(3) is equivalent to optimization problem:

$$\hat{\beta} = \arg\min_{\beta} |y - X\beta|^2, \quad s.t.\ (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \leq t \qquad (4)$$

- When $\alpha = 1$, the naive elastic net becomes ridge regression.
- Consider only $\alpha \in [0, 1)$,the elastic net penalty funciton is singular (without first derivation) at 0 and it is strictly convex for all $\alpha > 0$
- When $\alpha = 0$, the naive elastic net becomes lasso regression,it is convex but not strictly convex.

Solution of scenario (a)

This method minimizing equation(3) is equivalent to a lasso-type optimization problem.

- Given data set (**y**,**X**) and (λ_1,λ_2),define an artificial data set(**y***,**X***) by

$$\mathbf{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2}\begin{pmatrix}\mathbf{X}_{n\times p}\\ \sqrt{\lambda_2}\,\mathbf{I}_{p\times p}\end{pmatrix}, \quad \mathbf{y}^*_{(n+p)} = \begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix}$$

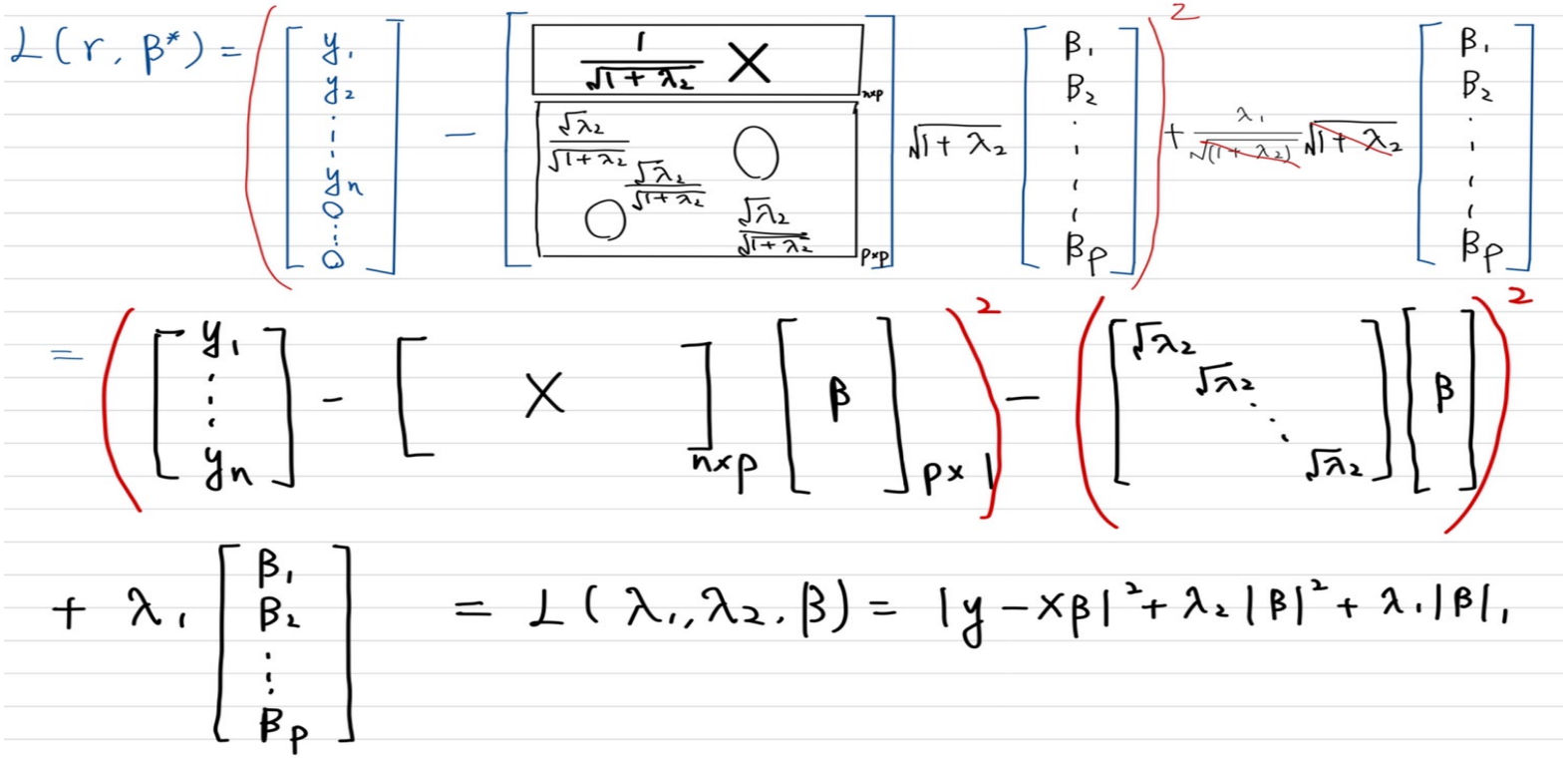
- Let $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$ and $\boldsymbol{\beta}^* = \sqrt{(1+\lambda_2)}\boldsymbol{\beta}$.Then the naive elastic net criterion can be written as

$$L(\gamma, \boldsymbol{\beta}^*) = |\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*|^2 + \gamma|\boldsymbol{\beta}^*|_1$$

- Let

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\beta} L\{(\gamma, \boldsymbol{\beta}^*)\} \quad \hat{\boldsymbol{\beta}} = \frac{1}{\sqrt{(1+\lambda_2)}}\hat{\boldsymbol{\beta}}^*$$

Solution of scenario (a)



- The sample size of $\mathbf{X}^*$ is n+p and x has rank p,which means the navie elastic net can potentially select all p predictors in all situations.
- This important property overcomes the limitations of the lasso in scenario (a).