03 - Lasso

ST 697 | Fall 2017
University of Alabama

Shrinkage Methods

Instead of an “all or nothing" approach, shrinkage methods force the coefficients closer toward 0.

- Usually this is accomplished through penalized regression where a penalty is imposed on the size of the coefficients
- Equivalently, the size of the coefficients are *constrained* not to exceed a threshold

The general framework is

$$\hat{\beta} = \arg\min_{\beta} \{l(\beta) + \lambda P(\beta)\}$$

where

- $l(\beta)$ is the loss function (e.g. mean squared error, negative log-likelihood)
- $\lambda \geq 0$ is the strength of the penalty
- $P(\beta)$ is the penalty term (as a function of the model parameters)

Two Representations

The penalized optimization (Lagrangian form)

$$\hat{\beta} = \arg\min_{\beta} \{l(\beta) + \lambda P(\beta)\}$$

An equivalent representation is (constrained optimization)

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta}\ l(\beta) \qquad \text{subject to } \ P(\beta) \leq t \\ &= \arg\min_{\beta:\ P(\beta)\leq t}\ l(\beta)\end{aligned}$$

Penalties

Examples penalties:

- Ridge Penalty

$$P(\beta) = \sum_{j=1}^{p} |\beta_j|^2 = \beta^{\mathsf{T}}\beta = \|\beta\|_2^2$$

- Lasso Penalty

$$P(\beta) = \sum_{j=1}^{p} |\beta_j| = \|\beta\|_1$$

- Best Subsets

$$P(\beta) = \sum_{j=1}^{p} |\beta_j|^0 = \sum_{j=1}^{p} 1_{(\beta_j \neq 0)}$$

The Lasso

For lasso regression

$$l(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$P(\beta) = \sum_{j=1}^{p} |\beta_j| \qquad \text{(Notice that } \beta_0 \text{ is not penalized)}$$

So the ridge solution becomes:

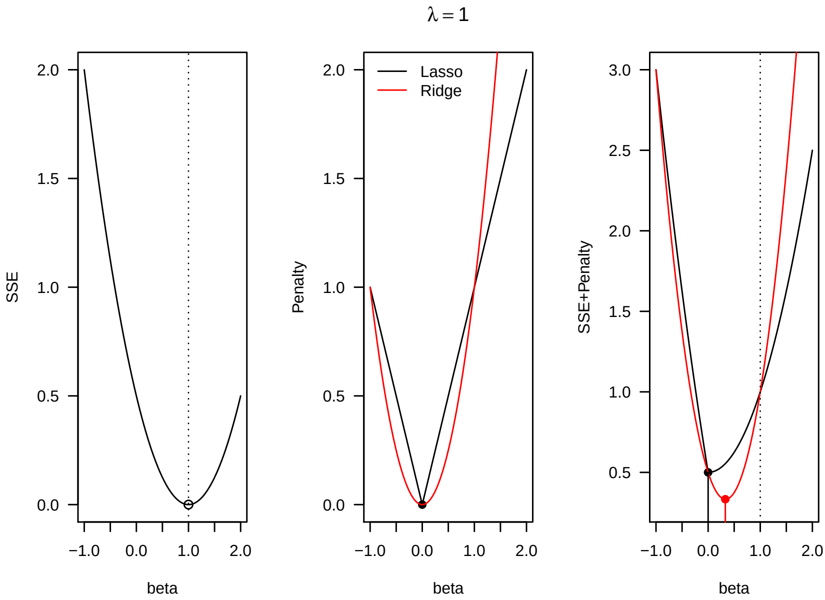
$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Why is it important to scale the predictor variables?

Lasso Penalty

- By using a L_1 penalty, lasso penalty can shrink some coefficients all the way to 0 (unlike the ridge penalty)
- This effectively removes predictors from the model (like the stepwise procedures), but in a type of continuous fashion
- Lasso stands for “Least Absolute Shrinkage and Selection Operator"

Lasso Selection: $l(\beta) = \frac{1}{2}(1-\beta)^2$



Geometry of LASSO and Ridge

$$\hat{\beta}^{pen} = \arg\min_{\beta} \; l(\beta) \qquad \text{subject to} \;\; P(\beta) \leq t$$

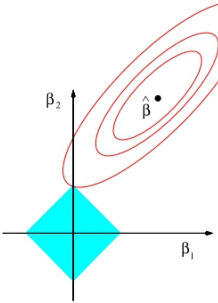


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Penalty Family

$$P(\beta, \alpha) = \sum_{j=1}^{p} |\beta_j|^q$$

- $q = 0$: Best subsets
- $q = 1$: Lasso
- $q = 2$: Ridge

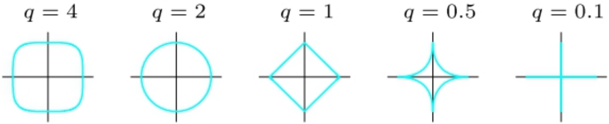
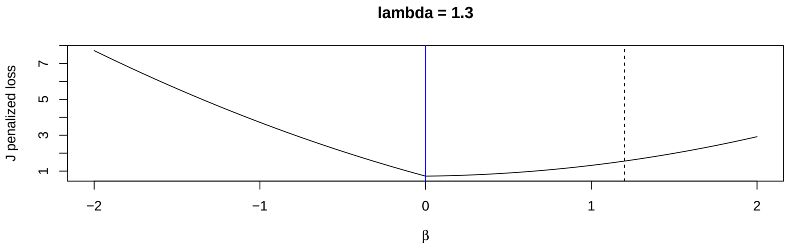
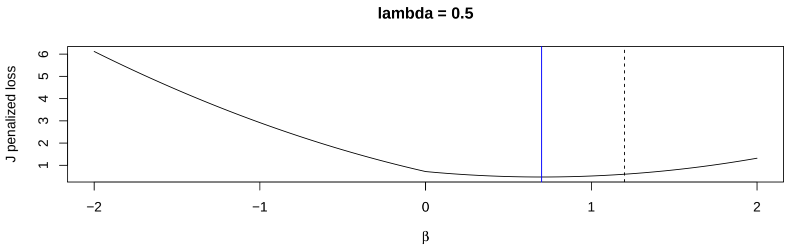


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Minimization function $J(\beta)$ for univariate lasso

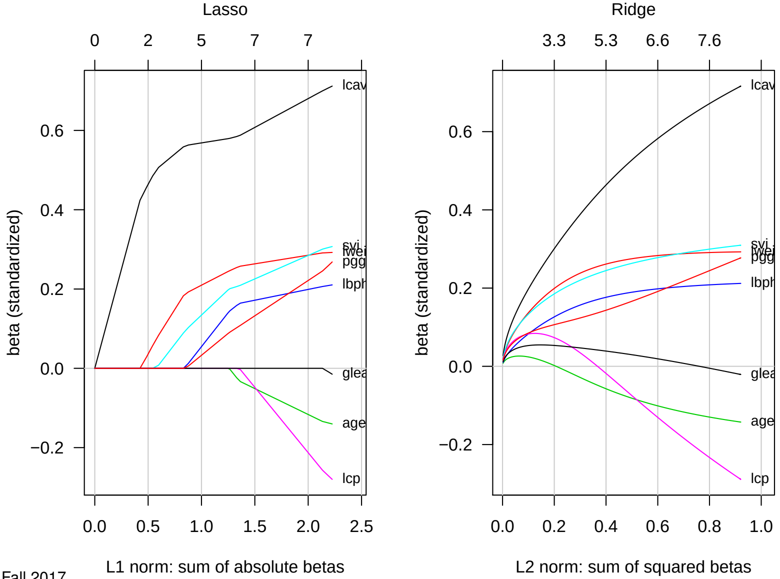
$$J(\beta, \lambda) = \frac{1}{2}(1.2 - \beta)^2 + \lambda\,|\beta|$$



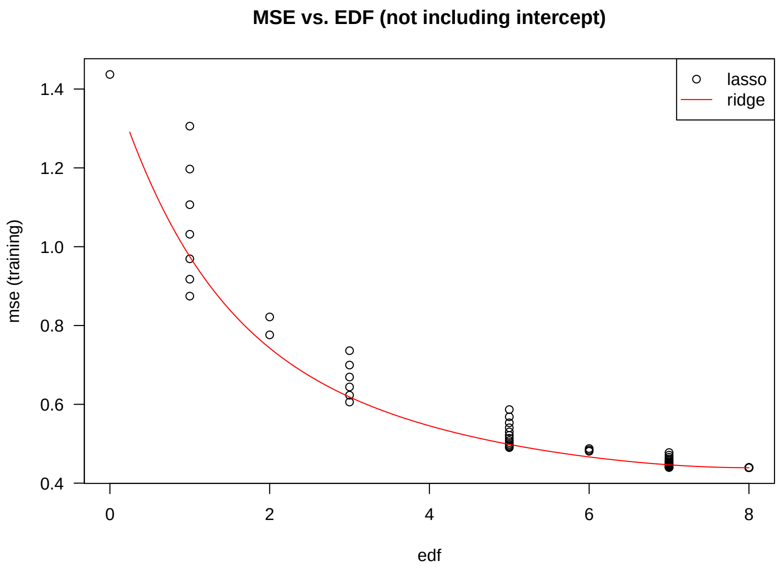


Comparing Lasso and Ridge Regression

Prostate Cancer Data from ESL book: Figs 3.8, 3.10 and Table 3.3

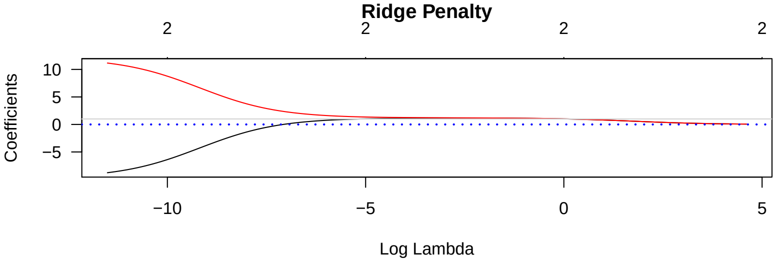


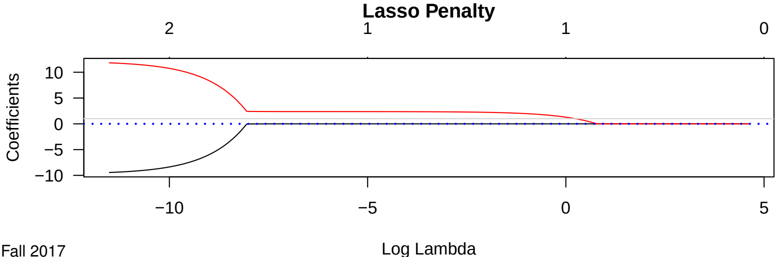
Comparing Lasso and Ridge Regression



Example with Strong Correlation

$$\boxed{Y = 1 + 1X_1 + 1X_2 + \epsilon}$$





Effective Number of Parameters

- Unlike ridge regression, the lasso is *not* a linear smoother. There is no way to write $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.
- Thus, estimating the effective degrees of freedom is not based on trace of hat matrix.
- It turns out that the number of non-zero coefficients is a decent approximation of the effective number of parameters
- We can use this value ($df = \sum_j \mathbb{1}(|\beta_j| > 0)$) in AIC/BIC/GCV for selecting λ
  - Note: the df is not continuous in λ, so the min SSE model would have smallest λ within the set with $df = k$

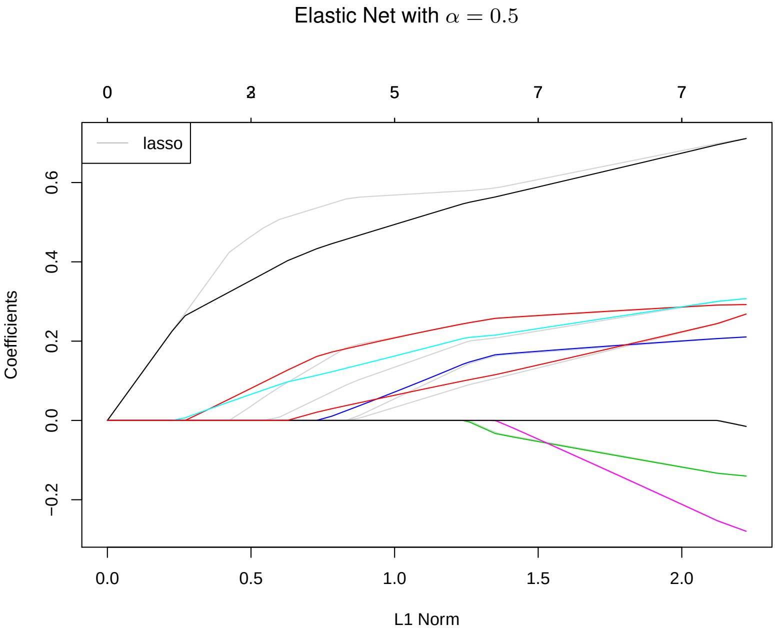
Elastic Net

The Elastic Net Penalty can help with selection (like lasso) and shrinks together correlated predictors (like ridge).

$$P(\beta, \alpha) = \sum_{j=1}^{p} \alpha \beta_j^2 + (1 - \alpha)|\beta_j| \qquad \text{Eq 3.54 on pg 73 of ESL}$$

$$P(\beta, \alpha) = \sum_{j=1}^{p} \frac{(1 - \alpha)}{2} \beta_j^2 + \alpha|\beta_j| \qquad \text{glmnet R package}$$

Compare Elastic Net to Lasso and Ridge



Categorical Predictors in Penalized Regression

1. How does lasso/ridge treat categorical predictors?
2. How does lasso/ridge treat interaction terms?
3. How does lasso/ridge treat basis expansions of a single variable, e.g. polynomial?

Group Lasso

- L groups of predictors
  - categorical variable with 3 levels will be in a group of 3 predictors
- Let X_l be $n \times p_l$ matrix of group l predictors
- β_l is $p_l \times 1$ group coefficients

$$J(\beta) = \ell(\beta) + P(\beta, \lambda)$$

$$\ell(\beta) = \left\| Y - \beta_0 1 - \sum_{l=1}^{L} X_l \beta_l \right\|_2^2$$

$$P(\beta, \lambda) = \sum_{l=1}^{L} \sqrt{p_l} \, \|\beta_l\|_2$$