30. Lasso Regression

November 24, 2025 9:42 AM

Last time (1970): Ridge Regression

Solve: $$\hat{\beta}^R_\lambda = \underset{\tilde{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \underbrace{\sum_{i=1}^n (Y_i - X_i^T \tilde{\beta})^2}_{\text{Squared Error}} + \lambda \underbrace{\sum_{j=1}^p \tilde{\beta}_j^2}_{\text{Quadratic penalty}} \right\}$$

- Closed form solution

$$\hat{\beta}^R_\lambda = (X^T X + \lambda I_p)^{-1} X^T Y$$
$$\hat{\beta} = (X^T X \quad)^{-1} X^T Y$$

- Shrinkage Estimator, send the estimator towards zero
  → Variance ↓ bias ↑

- Sometimes called L^2-regularization
- It Doesn't do variable selection → Hard to interpret

Lasso (1996) ⇒ Least absolute Selection + Shrinkage operator (Variable Selection, Ridge)

$$\hat{\beta}^L_\lambda = \underset{\tilde{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \underbrace{\sum_{i=1}^n (Y_i - X_i^T \tilde{\beta})^2}_{\text{Squared error}} + \lambda \underbrace{\sum_{j=1}^p |\tilde{\beta}_j|}_{\text{Absolute / linear penalty}} \right\}$$

(also called L^1-regularization)

- Why is it interesting?
  → Does Both Shrinkage + Selection at same time
  → Convex optimization Problem (Easier to solve than non-convex)
  → No closed form solution.

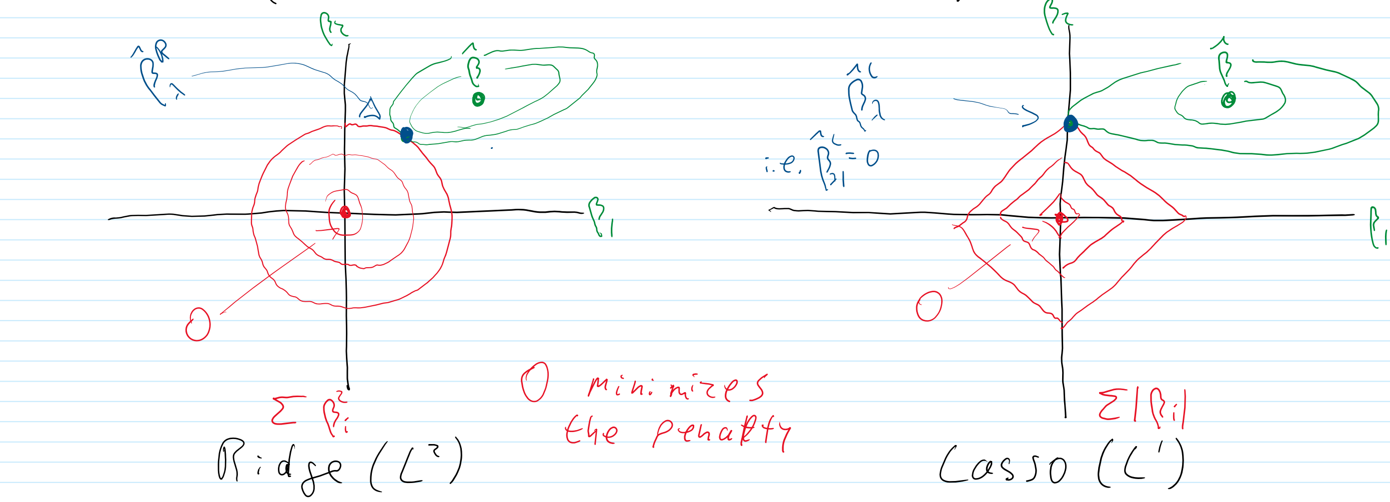
  * OLS and Ridge are unique in that Most Statistical Machine Learning Methods do not have a nice closed form solution ⇒ i.e. Need a computer!

  → This work spawned countless

Computer:

→ This work spawned countless Follow up papers on Theory.

Picture version of why Lasso sets some $\hat{\beta}^L_{\lambda j} = 0$ (i.e. does variable selection)



$\hat{\beta}$ Least squares so it is the "best" point to Minimize $\sum (Y_i - X_i^T \tilde{\beta})^2$

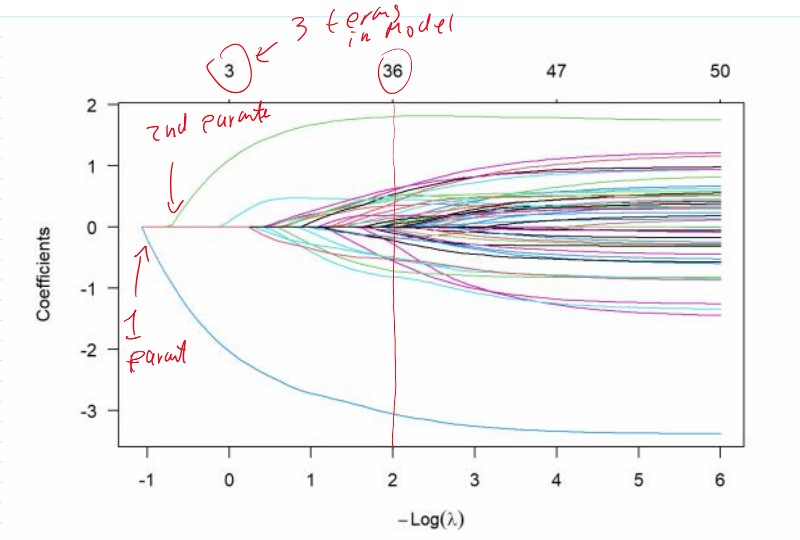
Elastic Net (2005): Zou + Hastie

why not do both Lasso + Ridge!

$$\hat{\beta}^{EN}_\lambda = \underset{\tilde{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \underbrace{\sum_{i=1}^n (Y_i - X_i^T \tilde{\beta})^2}_{\text{Squared Error}} + \lambda_1 \underbrace{\sum_{j=1}^p |\tilde{\beta}_j|}_{L^1} + \lambda_2 \underbrace{\sum_{j=1}^p \tilde{\beta}_j^2}_{L^2} \right\}$$

penalty

Claim: "Better" than Lasso but still does Shrinkage + Selection



"Lasso Paths"

- Each line is a parameter that "enters" the Model as $\lambda \downarrow$ or $-\log\lambda \uparrow$
- Every Choice of λ (cross section of plot) is a Lasso regression Model.

$\hat{\beta}^R_\lambda = (X^TX + \lambda I)^{-1} X^T Y$ or $\frac{\sum x_i y_i}{\sum x_i^2}$ or $\frac{\sum x_i y_i}{\sum x_i^2 + \lambda}$

$$\hat{\beta}^R_\lambda = (X^T X + \underbrace{\lambda I}) ^{-1} X^T Y \quad , \text{ e.g. } \quad \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{or} \quad \frac{\sum x_i y_i}{\sum x_i^2 + \lambda}$$

like adding λ to each term in the denominator

People have proven "sparsistency" → sparse + consistency

if the true β has, say, $\beta_i = 0$

then $\hat{\beta}^L_{\lambda\, i}$ will $= 0$ for large enough sample n.