PS5841

Data Science in Finance & Insurance

Shrinkage

Yubo Wang

Spring 2022

Ridge Regression

- Model

$$Y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

- Least squares: $\widehat{\boldsymbol{\beta}}^{LS}$ minimizes

$$R_{LS} = ESS_{LS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2$$

- Ridge regression: $\widehat{\boldsymbol{\beta}}_\lambda^R$ minimizes

$$R(\lambda) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

tuning parameter λ

no intercept

Solution

- Loss, where $\Lambda = \mathrm{diag}(0, \lambda_1, \dots, \lambda_p)$

$$R(\lambda) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \boldsymbol{\beta}^T\Lambda\boldsymbol{\beta}$$

$$\frac{\partial R}{\partial \boldsymbol{\beta}} = -2X^T(\boldsymbol{y} - X\boldsymbol{\beta}) + 2\Lambda\boldsymbol{\beta}$$
$$\widehat{\boldsymbol{\beta}}^R = (X^TX + \Lambda)^{-1}X^T\boldsymbol{y}$$

- Biased (when $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},\ var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$)

$$E\left(\widehat{\boldsymbol{\beta}}^R\right) = (X^TX + \Lambda)^{-1}X^TX\boldsymbol{\beta}$$
$$Var\left(\widehat{\boldsymbol{\beta}}^R\right) = \sigma^2(X^TX + \Lambda)^{-1}X^TX(X^TX + \Lambda)^{-1}$$

Equivalent Solution

- When features are centered, $\widehat{\boldsymbol{\beta}}_{\lambda}^{R}$ minimizes

no intercept

$$R(\lambda) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

$$\frac{\partial R}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T(\boldsymbol{y} - X\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}$$

$$\widehat{\boldsymbol{\beta}}^R = (X^TX + \lambda\boldsymbol{I})^{-1}X^T\boldsymbol{y}$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} y_i$$

Recipe

- Scale features
  - Centered and normalized
  - Standardized
- Estimate the intercept with OLS
  - The intercept is a measure of the mean of the response when features are zero
  - When features are centered, we have $\hat{\beta}_0 = \bar{y}$
- Estimate the remaining coefficients by a Ridge regression without intercept using the scaled features

Penalty Perspective

- Minimize

$$R(\lambda) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

- Equivalent formulation

$$\text{minimize } R(\lambda = 0) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})$$

$$\text{Subject to} \sum_{j=1}^{p} \beta_j^2 \leq s^2$$

- Tuning parameter λ
- Shrinks $\boldsymbol{\beta}$ and introduces bias

Bayesian Perspective

$$\boldsymbol{y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}), \qquad \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \tau^2 \boldsymbol{I})$$ prior

- Posterior log-likelihood

posterior likelihood prior

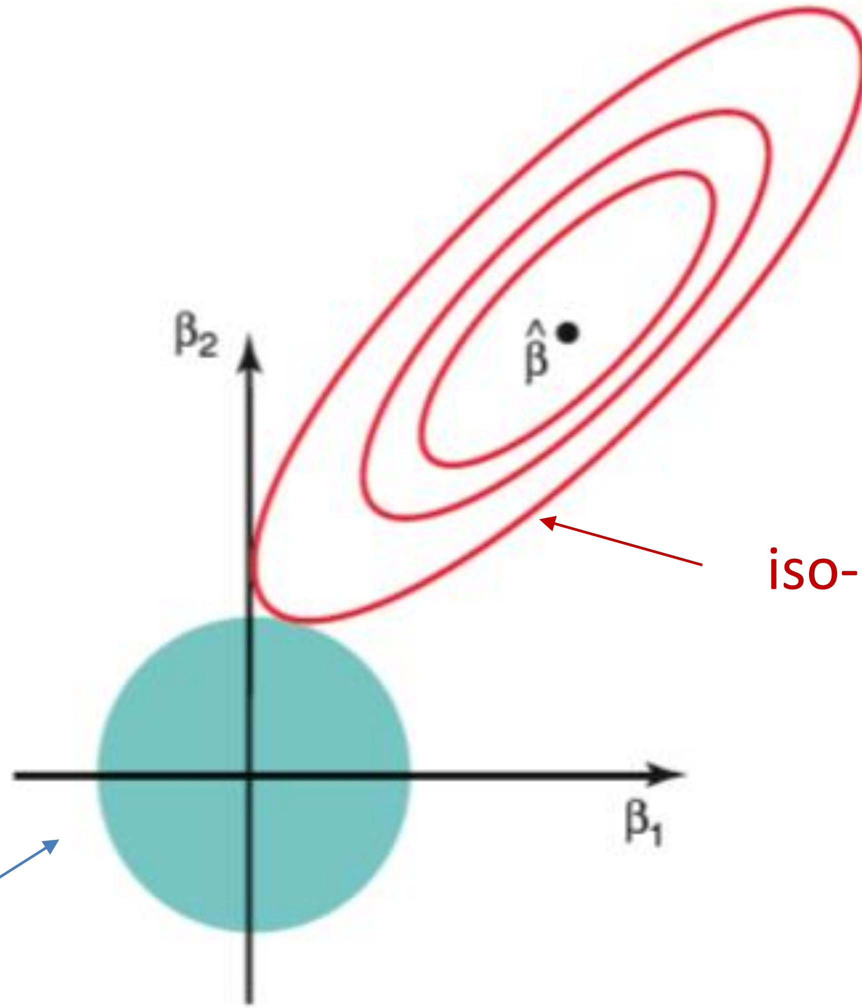
$$l(\boldsymbol{\beta};\,) \propto l(\boldsymbol{y};\,) + l(\boldsymbol{\beta})$$

$$= -\frac{1}{2\sigma^2}\left[(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \frac{\sigma^2}{\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right]$$

- Minimize $R(\lambda)$, $\lambda = \frac{\sigma^2}{\tau^2}$
- $\widehat{\boldsymbol{\beta}}^R$ maximizes the posterior

Geometry Perspective

Ridge estimate
$\beta_j \neq 0$



iso-ESS curve

$\beta_1^2 + \beta_2^2 \leq s^2$

LASSO Regression

- Model

$$Y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

- Least squares: $\widehat{\boldsymbol{\beta}}^{LS}$ minimizes

$$R_{LS} = ESS_{LS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2$$

- Ridge regression: $\widehat{\boldsymbol{\beta}}_\lambda^R$ minimizes

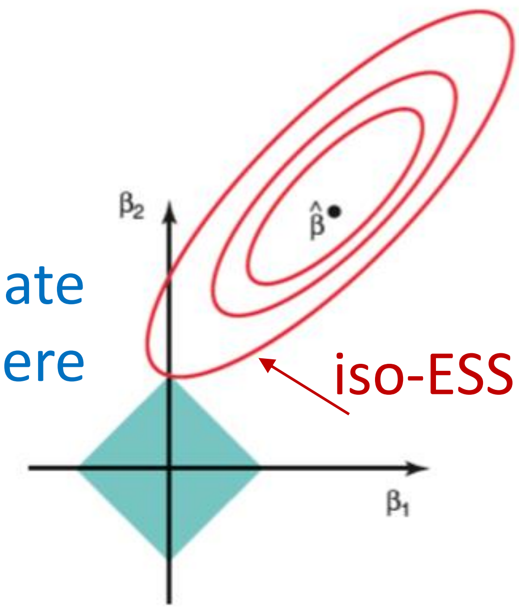
$$R(\lambda) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Generalized Penalty

- Minimize

$$ESS(\lambda) = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

- Ridge regression $q = 2$
- LASSO $q = 1$



LASSO estimate
$\beta_1 = 0$!!! here

iso-ESS curve

Bayesian Perspective

$$\boldsymbol{y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2\boldsymbol{I}), \qquad \boldsymbol{\beta} \sim Laplace(\boldsymbol{0}, 2\tau^2\boldsymbol{I})$$

prior

- Posterior log-likelihood

posterior likelihood prior

$$l(\boldsymbol{\beta};\,) \propto l(\boldsymbol{y};\,) + l(\boldsymbol{\beta})$$

$$= -\frac{1}{2\sigma^2}\left[(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \frac{\sigma^2}{\tau}|\boldsymbol{\beta}|\right]$$

- Minimize $R(\lambda)$, $\lambda = \frac{\sigma^2}{\tau}$
- $\widehat{\boldsymbol{\beta}}^L$ maximizes the posterior

Observations

- $\widehat{\boldsymbol{\beta}}^{LS} = \widehat{\boldsymbol{\beta}}^{R}_{\lambda=0} = \widehat{\boldsymbol{\beta}}^{L}_{\lambda=0}$
- The LASSO can perform variable selection since it can yield sparse models
- Ridge does better when the response is a function of many predictors with coefficients of roughly equal size
- LASSO does better when only a small portion of predictors have substantial coefficients

That was