**PS5841**

# Data Science in Finance & Insurance

## Regularization

Yubo Wang

Autumn 2022

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

# Regularization

- Variance reduction
- Loss with penalty

# Ridge Regression

- Model

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

- 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 $\lambda$

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

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

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

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

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

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

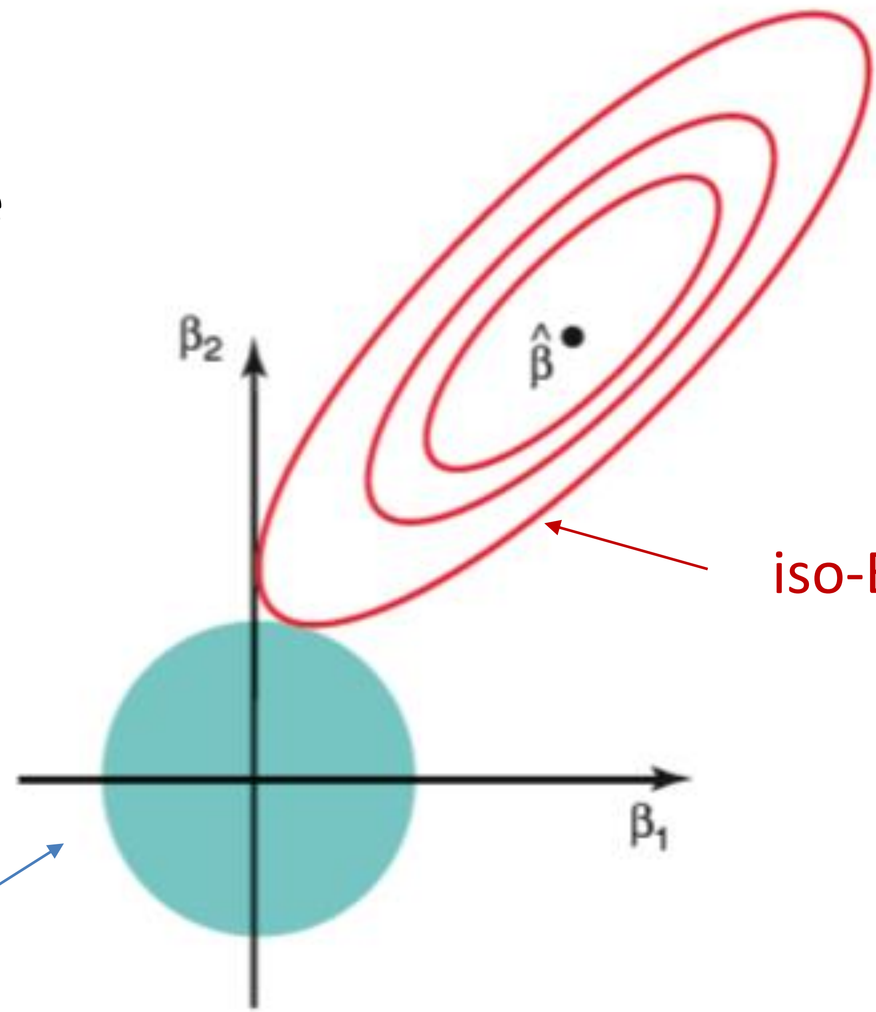
# Penalty Perspective

- 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 $\lambda$
- Shrinks $\boldsymbol{\beta}$ and introduces bias

# Geometry Perspective

Ridge estimate

$\beta_j \neq 0$




iso-ESS curve

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

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

# LASSO

- Model

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

- Least Absolute Shrinkage and Selection Operator (LASSO) $\widehat{\boldsymbol{\beta}}_\lambda^L$ minimizes

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

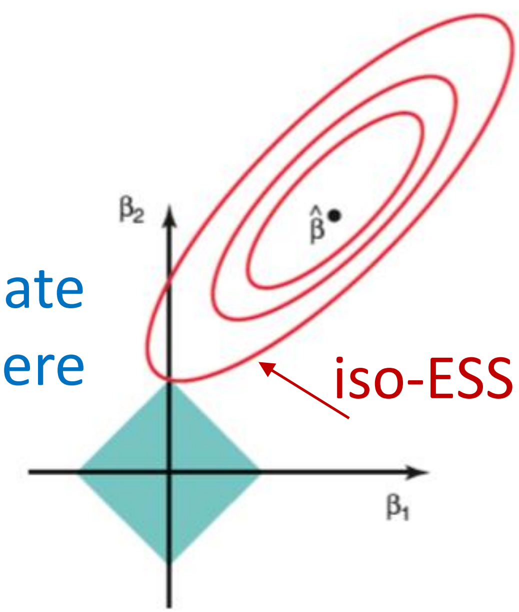
tuning parameter $\lambda$ no intercept

# Penalty & Geometry Perspective

- Minimize

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

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




iso-ESS curve

# Bayesian Perspective

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

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

# Decision Tree Pruning

- A “big” tree risks overfitting data

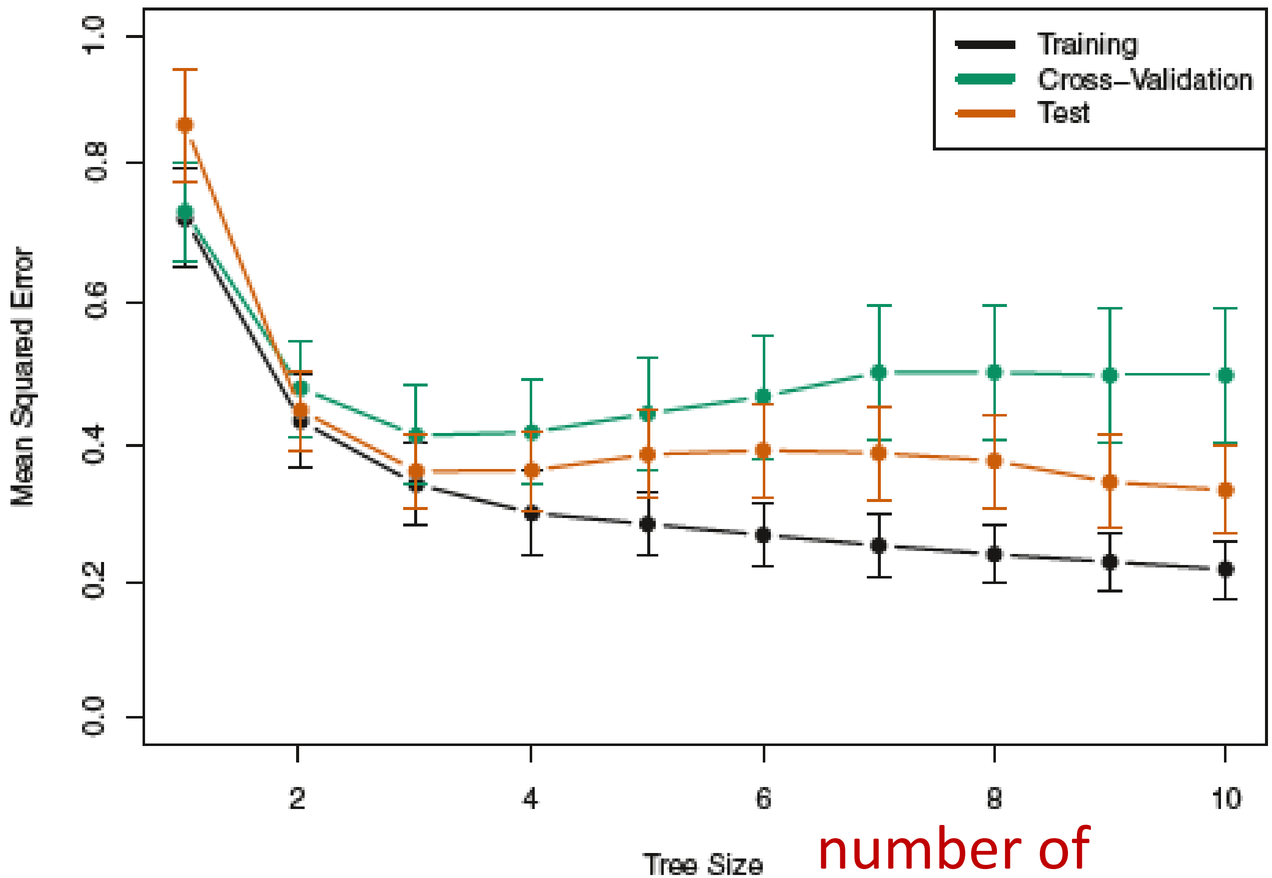
# Cost Complexity Pruning

- Loss

$$\text{error\_rate} + \alpha|T|$$
$$MSE + \alpha|T|$$

- tuning parameter $(\alpha \geq 0)$
- Resulting tree is a subtree $T \subset T_0$ which minimizes the loss

# Example: Pruning




number of terminal nodes

# That was