Political Methodology III: Model Based Inference

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

May 16th, 2017

Model Based Inference

1) Likelihood inference
2) Machine Learning
    a) Model Selection
    b) Unsupervised Latent Features
    c) Classification/Prediction/Regression

Supervised Learning Methods

1) Task
    - Classify objects into pre existing categories
    - Predict some future outcome
    - Learn a response surface for causal inference
2) Objective function
    1) Penalized Regressions
        - Ridge regression
        - LASSO regression
3) Optimization
    1) Ridge regression $\rightsquigarrow$ Straightforward modification of OLS
    2) LASSO regression $\rightsquigarrow$ Coordinate Descent
4) Validation
    - Obtain predicted fit for new data $f(\boldsymbol{X}_i, \widehat{\boldsymbol{\theta}})$
    - Examine prediction performance $\rightsquigarrow$ compare prediction/classification to gold standard

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow \{\text{not, credit claiming}\}$

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow$ {not, credit claiming}

We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow$ {not, credit claiming}
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i \right)^2$$

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \leadsto$ {not, credit claiming}
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \\
\widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \right\}
\end{aligned}
$$

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow$ {not, credit claiming}
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \\ \widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \right\} \\ &= \left(\boldsymbol{X}^{'} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{'} \boldsymbol{Y} \end{aligned}$$

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \leadsto$ {not, credit claiming}
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \\ \widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \right\} \\ &= \left(\boldsymbol{X}^{'} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{'} \boldsymbol{Y} \end{aligned}$$

Problem:

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow$ {not, credit claiming}
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'}\boldsymbol{x}_i\right)^2 \\ \widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'}\boldsymbol{x}_i\right)^2 \right\} \\ &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{'}\boldsymbol{Y} \end{aligned}$$

Problem:

- J will likely be large (perhaps $J > N$)

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow$ {not, credit claiming}
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \\
\widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \right\} \\
&= \left(\boldsymbol{X}^{'} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{'} \boldsymbol{Y}
\end{aligned}
$$

Problem:

- J will likely be large (perhaps $J > N$)
- There many correlated variables

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow$ {not, credit claiming}
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \\
\widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \right\} \\
&= \left(\boldsymbol{X}^{'} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{'} \boldsymbol{Y}
\end{aligned}
$$

Problem:

- J will likely be large (perhaps $J > N$)
- There many correlated variables

Predictions will be variable

Mean Square Error

Suppose θ is some value of the true parameter

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\mathsf{E}[(\hat{\theta} - \theta)^2]$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\mathsf{E}[(\hat{\theta} - \theta)^2] = E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned}
\mathsf{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2
\end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned}
\mathsf{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + (E[\widehat{\theta} - \theta])^2
\end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} \quad = \quad E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned}
\mathsf{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + (E[\widehat{\theta} - \theta])^2 \\
&= \text{Var}(\theta) + \text{Bias}^2
\end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned}
\text{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + (E[\widehat{\theta} - \theta])^2 \\
&= \text{Var}(\theta) + \text{Bias}^2
\end{aligned}$$

To reduce MSE, we are willing to induce bias to decrease variance $\rightsquigarrow$ methods that shrink coefficeints toward zero

Ridge Regression

Penalty for model complexity

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y})$$

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \sum_{i=1}^{N} \left(y_i - \ \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2$$

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \underbrace{\lambda\sum_{j=1}^{J}\beta_j^2}_{\text{Penalty}}$$

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \;=\; \sum_{i=1}^{N} \left(y_i - \;\beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

where:

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \sum_{i=1}^{N} \left(y_i - \ \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

where:

- $\beta_0 \rightsquigarrow$ intercept

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \sum_{i=1}^{N} \left(y_i - \ \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

where:

- $\beta_0 \rightsquigarrow$ intercept
- $\lambda \rightsquigarrow$ penalty parameter

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \sum_{i=1}^{N} \left(y_i - \ \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

where:

- $\beta_0 \rightsquigarrow$ intercept
- $\lambda \rightsquigarrow$ penalty parameter
- Standardized $\boldsymbol{X}$ (coefficients on same scale)

Ridge Regression$\rightsquigarrow$ Optimization

$$\boldsymbol{\beta}^{\text{Ridge}} \quad = \quad \text{arg min}_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \right\}$$

Ridge Regression⇝ Optimization

$$\begin{aligned}\boldsymbol{\beta}^{\textsf{Ridge}} &= \textsf{arg min}_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \right\} \\ &= \textsf{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \right\}\end{aligned}$$

Ridge Regression$\rightsquigarrow$ Optimization

$$
\begin{aligned}
\boldsymbol{\beta}^{\text{Ridge}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta})^{'}(\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{'}\boldsymbol{\beta} \right\}
\end{aligned}
$$

Demean the data and set $\beta_0 = \bar{y} = \sum_{i=1}^{N} \frac{y_i}{N}$

Ridge Regression$\rightsquigarrow$ Optimization

$$
\begin{aligned}
\boldsymbol{\beta}^{\text{Ridge}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta})^{'}(\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{'}\boldsymbol{\beta} \right\} \\
&= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J \right)^{-1} \boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}
$$

Demean the data and set $\beta_0 = \bar{y} = \sum_{i=1}^{N} \frac{y_i}{N}$

Ridge Regression$\rightsquigarrow$ Optimization

$$\begin{aligned}
\boldsymbol{\beta}^{\textsf{Ridge}} &= \textsf{arg min}_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \right\} \\
&= \textsf{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \right\} \\
&= \textsf{arg min}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta})^{'}(\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{'}\boldsymbol{\beta} \right\} \\
&= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J \right)^{-1} \boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}$$

Demean the data and set $\beta_0 = \bar{y} = \sum_{i=1}^{N} \frac{y_i}{N}$

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}$$

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\ &= \boldsymbol{X}^{'}\boldsymbol{Y}\end{aligned}$$

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\text{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}$$

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\textsf{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}$$

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\textsf{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)^{-1}\widehat{\boldsymbol{\beta}}
\end{aligned}$$

Ridge Regression⇝ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\textsf{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)^{-1}\widehat{\boldsymbol{\beta}} \\
\beta_j^{\textsf{Ridge}} &= \frac{\widehat{\beta}_j}{1+\lambda}
\end{aligned}
$$

Ridge Regression$\rightsquigarrow$ Intuition (2)

$$\begin{aligned}\boldsymbol{\beta}_j & \sim \textsf{Normal}(0, \tau^2) \\ y_i & \sim \textsf{Normal}(\beta_0 + \boldsymbol{x}_i^{'}\boldsymbol{\beta}, \sigma^2)\end{aligned}$$

Ridge Regression⇝ Intuition (2)

$$\begin{aligned}\boldsymbol{\beta}_j & \sim \textsf{Normal}(0, \tau^2) \\ y_i & \sim \textsf{Normal}(\beta_0 + \boldsymbol{x}_i^{'}\boldsymbol{\beta}, \sigma^2)\end{aligned}$$

$$p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) \quad \propto \quad \prod_{j=1}^{J} p(\beta_j) \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})$$

Ridge Regression$\leadsto$ Intuition (2)

$$\begin{aligned}\boldsymbol{\beta}_j &\sim \textsf{Normal}(0, \tau^2)\\ y_i &\sim \textsf{Normal}(\beta_0 + \boldsymbol{x}_i^{'}\boldsymbol{\beta}, \sigma^2)\end{aligned}$$

$$\begin{aligned}p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) &\propto \prod_{j=1}^{J} p(\beta_j) \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})\\ &\propto \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{\beta_j^2}{2\tau^2}\right) \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta_0 - \boldsymbol{x}_i^{'}\boldsymbol{\beta})^2}{2\sigma^2}\right)\end{aligned}$$

Ridge Regression⇝ Intuition (2)

$$\log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) \quad = \quad -\sum_{j=1}^{J} \frac{\beta_j^2}{2\tau^2} - \sum_{i=1}^{N} \frac{(y_i - \beta_0 - \boldsymbol{x}'\boldsymbol{\beta})^2}{2\sigma^2}$$

Ridge Regression⇝ Intuition (2)

$$\begin{aligned}
\log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) &= -\sum_{j=1}^{J} \frac{\beta_j^2}{2\tau^2} - \sum_{i=1}^{N} \frac{(y_i - \beta_0 - \boldsymbol{x}^{'}\boldsymbol{\beta})^2}{2\sigma^2} \\
-2\sigma^2 \log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N}(y_i - \beta_0 - \boldsymbol{x}^{'}\boldsymbol{\beta})^2 + \sum_{j=1}^{J} \frac{\sigma^2}{\tau^2}\beta_j^2
\end{aligned}$$

Ridge Regression$\rightsquigarrow$ Intuition (2)

$$\begin{aligned}
\log p(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y}) &= -\sum_{j=1}^{J}\frac{\beta_j^2}{2\tau^2} - \sum_{i=1}^{N}\frac{(y_i - \beta_0 - \boldsymbol{x}^{'}\boldsymbol{\beta})^2}{2\sigma^2} \\
-2\sigma^2 \log p(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y}) &= \sum_{i=1}^{N}(y_i - \beta_0 - \boldsymbol{x}^{'}\boldsymbol{\beta})^2 + \sum_{j=1}^{J}\frac{\sigma^2}{\tau^2}\beta_j^2
\end{aligned}$$

where:

Ridge Regression⇝ Intuition (2)

$$\begin{aligned}
\log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) &= -\sum_{j=1}^{J} \frac{\beta_j^2}{2\tau^2} - \sum_{i=1}^{N} \frac{(y_i - \beta_0 - \boldsymbol{x}^{'}\boldsymbol{\beta})^2}{2\sigma^2} \\
-2\sigma^2 \log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N}(y_i - \beta_0 - \boldsymbol{x}^{'}\boldsymbol{\beta})^2 + \sum_{j=1}^{J} \frac{\sigma^2}{\tau^2}\beta_j^2
\end{aligned}$$

where:

- $\lambda = \frac{\sigma^2}{\tau^2}$

Ridge Regression ⇝ Intuition (3)

Definition

Suppose $\boldsymbol{X}$ is an $N \times J$ matrix. Then $\boldsymbol{X}$ can be written as:

$$\boldsymbol{X} = \underbrace{\boldsymbol{U}}_{N\times N}\underbrace{\boldsymbol{S}}_{N\times J}\underbrace{\boldsymbol{V}^{'}}_{J\times J}$$

Where:

$$\begin{aligned}\boldsymbol{U}^{'}\boldsymbol{U} &= \boldsymbol{I}_N \\ \boldsymbol{V}^{'}\boldsymbol{V} &= \boldsymbol{V}\boldsymbol{V}^{'} = \boldsymbol{I}_J\end{aligned}$$

$\boldsymbol{S}$ contains $\min(N, J)$ singular values, $\sqrt{\lambda_j} \geq 0$ down the diagonal and then 0's for the remaining entries

Ridge Regression $\rightsquigarrow$ Intuition (3)

Recall: PCA:

$$\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X} = \underbrace{\boldsymbol{W}}_{\text{eigenvectors}} \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_J \end{pmatrix} \underbrace{\boldsymbol{W}^{'}}_{\text{eigenvectors}}$$

Ridge Regression ⇝ Intuition (3)

Recall: PCA:

$$\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X} = \underbrace{\boldsymbol{W}}_{\text{eigenvectors}} \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_J \end{pmatrix} \underbrace{\boldsymbol{W}^{'}}_{\text{eigenvectors}}$$

Using SVD:

Ridge Regression $\rightsquigarrow$ Intuition (3)

Recall: PCA:

$$\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X} = \underbrace{\boldsymbol{W}}_{\text{eigenvectors}} \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_J \end{pmatrix} \underbrace{\boldsymbol{W}^{'}}_{\text{eigenvectors}}$$

Using SVD:

$$\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X} \quad = \quad \boldsymbol{V}\boldsymbol{S}^{'} \underbrace{\left(\boldsymbol{U}^{'}\boldsymbol{U}\right)}_{\boldsymbol{I}_J} \boldsymbol{S}\boldsymbol{V}^{'}$$

Ridge Regression ⇝ Intuition (3)

Recall: PCA:

$$\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X} = \underbrace{\boldsymbol{W}}_{\text{eigenvectors}} \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_J \end{pmatrix} \underbrace{\boldsymbol{W}^{'}}_{\text{eigenvectors}}$$

Using SVD:

$$\begin{aligned} \frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X} &= \boldsymbol{V}\boldsymbol{S}^{'}\underbrace{\left(\boldsymbol{U}^{'}\boldsymbol{U}\right)}_{\boldsymbol{I}_J}\boldsymbol{S}\boldsymbol{V}^{'} \\ &= \frac{1}{N}\boldsymbol{V}\boldsymbol{S}^{'}\boldsymbol{S}\boldsymbol{V}^{'} \end{aligned}$$

Ridge Regression $\rightsquigarrow$ Intuition (3)

Recall: PCA:

$$\frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X} = \underbrace{\boldsymbol{W}}_{\text{eigenvectors}} \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_J \end{pmatrix} \underbrace{\boldsymbol{W}^{'}}_{\text{eigenvectors}}$$

Using SVD:

$$\begin{aligned} \frac{1}{N}\boldsymbol{X}^{'}\boldsymbol{X} &= \boldsymbol{V}\boldsymbol{S}^{'}\underbrace{\left(\boldsymbol{U}^{'}\boldsymbol{U}\right)}_{\boldsymbol{I}_J}\boldsymbol{S}\boldsymbol{V}^{'} \\ &= \frac{1}{N}\boldsymbol{V}\boldsymbol{S}^{'}\boldsymbol{S}\boldsymbol{V}^{'} \\ &= \underbrace{\boldsymbol{V}}_{\text{eigenvectors}} \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_J \end{pmatrix} \underbrace{\boldsymbol{V}^{'}}_{\text{eigenvectors}} \end{aligned}$$

Ridge Regression ⇝ Intuition (3)

We can write the predicted values for a regular regression as

$$\begin{aligned}
\hat{Y} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{U}\boldsymbol{U}^{'}\boldsymbol{Y} = \sum_{j=1}^{J}\boldsymbol{u}_j\boldsymbol{u}_j^{'}\boldsymbol{Y}
\end{aligned}$$

Ridge Regression $\rightsquigarrow$ Intuition (3)

We can write the predicted values for a regular regression as

$$\begin{aligned}
\hat{Y} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{U}\boldsymbol{U}^{'}\boldsymbol{Y} = \sum_{j=1}^{J}\boldsymbol{u}_j\boldsymbol{u}_j^{'}\boldsymbol{Y}
\end{aligned}$$

We can write $\boldsymbol{\beta}^{\text{ridge}}$ as

Ridge Regression $\rightsquigarrow$ Intuition (3)

We can write the predicted values for a regular regression as

$$\begin{aligned}
\hat{Y} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{U}\boldsymbol{U}^{'}\boldsymbol{Y} = \sum_{j=1}^{J}\boldsymbol{u}_j\boldsymbol{u}_j^{'}\boldsymbol{Y}
\end{aligned}$$

We can write $\boldsymbol{\beta}^{\text{ridge}}$ as

$$\hat{Y}^{\text{ridge}} = \boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}$$

Ridge Regression ⇝ Intuition (3)

We can write the predicted values for a regular regression as

$$\begin{aligned}\hat{Y} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\ &= \boldsymbol{U}\boldsymbol{U}^{'}\boldsymbol{Y} = \sum_{j=1}^{J}\boldsymbol{u}_j\boldsymbol{u}_j^{'}\boldsymbol{Y}\end{aligned}$$

We can write $\boldsymbol{\beta}^{\text{ridge}}$ as

$$\begin{aligned}\hat{Y}^{\text{ridge}} &= \boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\ &= \boldsymbol{U}\tilde{\boldsymbol{S}}\boldsymbol{U}^{'}\boldsymbol{Y}\end{aligned}$$

Where

$$\tilde{\boldsymbol{S}} = \left[\boldsymbol{S}(\boldsymbol{S}^{'}\boldsymbol{S} + \lambda\boldsymbol{I}_J)^{-1}\boldsymbol{S}\right]$$

Ridge Regression $\rightsquigarrow$ Intuition (3)

We can write the predicted values for a regular regression as

$$\begin{aligned}\hat{Y} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\ &= \boldsymbol{U}\boldsymbol{U}^{'}\boldsymbol{Y} = \sum_{j=1}^{J}\boldsymbol{u}_j\boldsymbol{u}_j^{'}\boldsymbol{Y}\end{aligned}$$

We can write $\boldsymbol{\beta}^{\text{ridge}}$ as

$$\begin{aligned}\hat{Y}^{\text{ridge}} &= \boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\ &= \boldsymbol{U}\tilde{\boldsymbol{S}}\boldsymbol{U}^{'}\boldsymbol{Y}\end{aligned}$$

Where

$$\tilde{\boldsymbol{S}} = \left[\boldsymbol{S}(\boldsymbol{S}^{'}\boldsymbol{S} + \lambda\boldsymbol{I}_J)^{-1}\boldsymbol{S}\right]$$

Which we can write as:

$$\hat{Y}^{\text{ridge}} = \sum_{j=1}^{J}\boldsymbol{u}_j\frac{\lambda_j}{\lambda_j + \lambda}\boldsymbol{u}_j^{'}\boldsymbol{Y}$$

Degrees of Freedom for Ridge

We will say that the degrees of freedom for Ridge regression with penalty λ is

$$\text{dof}(\lambda) \quad = \quad \sum_{j=1}^{J} \frac{\lambda_j}{\lambda_j + \lambda}$$

Lasso Regression Objective Function

Different Penalty for Model Complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

Lasso Regression Objective Function

Different Penalty for Model Complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

Lasso Regression Optimization

Definition

Coordinate Descent Algorithms:
Consider $g : \Re^J \rightarrow \Re$. Our goal is to find $\boldsymbol{x}^* \in \Re^J$ such that $g(\boldsymbol{x}^*) \leq g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Re$.
To find $\boldsymbol{x}^*$:
Until convergence: for each iteration t and each coordinate j

$$x_j^{t+1} = \text{arg min}_{x_j \in \Re} g(x_1^{t+1}, x_2^{t+1}, \ldots, x_{j-1}^{t+1}, x_j, x_{j+1}^t, \ldots, x_J^t)$$

Lasso Regression Optimization: Coordinate Descent

$$\tilde{f}(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{J}|\beta_j|$$

Lasso Regression Optimization: Coordinate Descent

$$\tilde{f}(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{J}|\beta_j|$$

- Case 1: If $\beta_j = 0 \rightsquigarrow$ not differentiable. But $\beta_j = 0$

Lasso Regression Optimization: Coordinate Descent

$$\tilde{f}(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \;=\; \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{J}|\beta_j|$$

- Case 1: If $\beta_j = 0 \rightsquigarrow$ not differentiable. But $\beta_j = 0$
- Case 2: If $\beta_j > (<) 0 \rightsquigarrow$ differentiable $\rightsquigarrow$ differentiate and solve for β_j

Lasso Regression Optimization: Coordinate Descent

$$\tilde{f}(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{J}|\beta_j|$$

- Case 1: If $\beta_j = 0 \rightsquigarrow$ not differentiable. But $\beta_j = 0$
- Case 2: If $\beta_j > (<)0 \rightsquigarrow$ differentiable $\rightsquigarrow$ differentiate and solve for β_j

Define $\tilde{y}_i^j = \beta_0 + \sum_{l \neq j} x_{il}\beta_l$

Lasso Regression Optimization: Coordinate Descent

$$\tilde{f}(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} |\beta_j|$$

- Case 1: If $\beta_j = 0 \rightsquigarrow$ not differentiable. But $\beta_j = 0$
- Case 2: If $\beta_j > (<) 0 \rightsquigarrow$ differentiable $\rightsquigarrow$ differentiate and solve for β_j

Define $\tilde{y}_i^j = \beta_0 + \sum_{l \neq j} x_{il} \beta_l$
$r^j \equiv \frac{1}{N} \sum_{i=1}^{N} x_{ij} (y_i - \tilde{y}_i^j)$

Lasso Regression Optimization: Coordinate Descent

$$\tilde{f}(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} |\beta_j|$$

- Case 1: If $\beta_j = 0 \rightsquigarrow$ not differentiable. But $\beta_j = 0$
- Case 2: If $\beta_j > (<) 0 \rightsquigarrow$ differentiable $\rightsquigarrow$ differentiate and solve for β_j

Define $\tilde{y}_i^j = \beta_0 + \sum_{l \neq j} x_{il} \beta_l$
$r^j \equiv \frac{1}{N} \sum_{i=1}^{N} x_{ij} (y_i - \tilde{y}_i^j)$
Update step for β_j is

Lasso Regression Optimization: Coordinate Descent

$$\tilde{f}(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{J}|\beta_j|$$

- Case 1: If $\beta_j = 0 \rightsquigarrow$ not differentiable. But $\beta_j = 0$
- Case 2: If $\beta_j > (<)0 \rightsquigarrow$ differentiable $\rightsquigarrow$ differentiate and solve for β_j

Define $\tilde{y}_i^j = \beta_0 + \sum_{l \neq j} x_{il}\beta_l$
$r^j \equiv \frac{1}{N}\sum_{i=1}^{N} x_{ij}(y_i - \tilde{y}_i^j)$
Update step for β_j is

$$\beta_j \quad \leftarrow \quad \text{sign}(r^j)\text{max}(|r^j| - \lambda, 0)$$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j|$$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j| \\
&= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j|
\end{aligned}
$$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j| \\
&= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j|
\end{aligned}
$$

The coefficient is

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j| \\ &= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j| \end{aligned}$$

The coefficient is

$$\beta_j^{\textsf{LASSO}} = \textsf{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_{+}$$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j| \\
&= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j|
\end{aligned}
$$

The coefficient is

$$
\beta_j^{\mathsf{LASSO}} = \mathsf{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_{+}
$$

- $\mathsf{sign}(\cdot) \rightsquigarrow 1$ or -1

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j| \\ &= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j| \end{aligned}$$

The coefficient is

$$\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_{+}$$

- $\text{sign}(\cdot) \rightsquigarrow 1$ or -1
- $\left(|\widehat{\beta}_j| - \lambda\right)_{+} = \text{max}(|\widehat{\beta}_j| - \lambda, 0)$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

Lasso Regression⇝ Intuition 1, Soft Thresholding

Compare soft assignment

$$\beta_j^{\mathsf{LASSO}} = \mathsf{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

$$\beta_j^{\mathsf{LASSO}} \quad = \quad \mathsf{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

With hard assignment, selecting M biggest components

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

$$\beta_j^{\textsf{LASSO}} = \textsf{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

With hard assignment, selecting M biggest components

$$\beta_j^{\textsf{subset}} = \widehat{\beta}_j \cdot I\left(|\widehat{\beta}_j| \geq |\widehat{\beta}_{(M)}|\right)$$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

$$\beta_j^{\textsf{LASSO}} \quad = \quad \textsf{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

With hard assignment, selecting M biggest components

$$\beta_j^{\textsf{subset}} \quad = \quad \widehat{\beta}_j \cdot I\left(|\widehat{\beta}_j| \geq |\widehat{\beta}_{(M)}|\right)$$

Intuition 2: Prior on coefficients $\rightsquigarrow$ Laplace "The Bayesian LASSO"

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

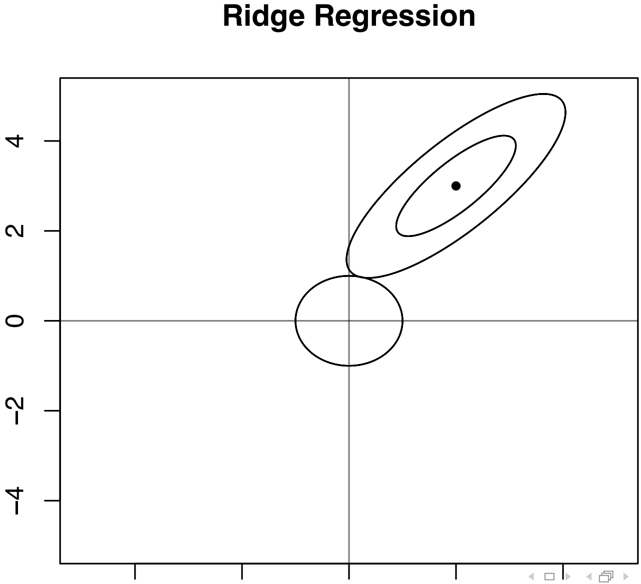
$$\beta_j^{\textsf{LASSO}} \quad = \quad \textsf{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

With hard assignment, selecting M biggest components

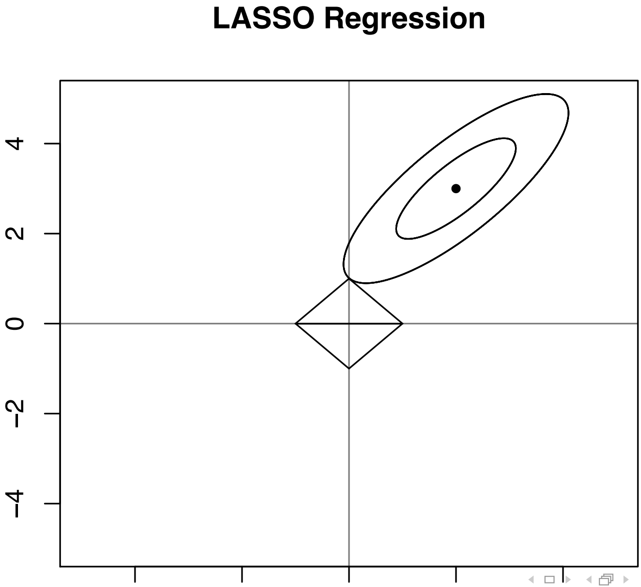
$$\beta_j^{\textsf{subset}} \quad = \quad \widehat{\beta}_j \cdot I\left(|\widehat{\beta}_j| \geq |\widehat{\beta}_{(M)}|\right)$$

Intuition 2: Prior on coefficients $\rightsquigarrow$ Laplace "The Bayesian LASSO"
Why does LASSO induce sparsity?

Comparing Ridge and LASSO



Comparing Ridge and LASSO



Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\sum_{j=1}^{2} \beta_j^2 \quad = \quad \frac{1}{2} + \frac{1}{2} = 1$$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned}
\sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\
\sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1
\end{aligned}$$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned}
\sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\
\sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1
\end{aligned}$$

Under LASSO

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned} \sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\ \sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1 \end{aligned}$$

Under LASSO

$$\sum_{j=1}^{2} |\beta_j| = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned}\sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\ \sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1\end{aligned}$$

Under LASSO

$$\begin{aligned}\sum_{j=1}^{2} |\beta_j| &= \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2} \\ \sum_{j=1}^{2} |\tilde{\beta}_j| &= 1 + 0 = 1\end{aligned}$$

Ridge and LASSO: The Elastic-Net

Combining the two criteria $\rightsquigarrow$ Elastic-Net

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right)$$

Ridge and LASSO: The Elastic-Net

Combining the two criteria $\rightsquigarrow$ Elastic-Net

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right)$$

The new update step (for coordinate descent:)

Ridge and LASSO: The Elastic-Net

Combining the two criteria $\rightsquigarrow$ Elastic-Net

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{J}\left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right)$$

The new update step (for coordinate descent:)

$$\beta_j \quad \leftarrow \quad \frac{\textsf{sign}(r^j)\textsf{max}(|r^j| - \lambda\alpha, 0)}{1 + \lambda(1-\alpha)}$$

Selecting λ

How do we determine λ? $\rightsquigarrow$ Cross validation (Recall code from Monday)

Selecting λ

How do we determine λ? $\rightsquigarrow$ Cross validation (Recall code from Monday)
Applying models gives score (probability) of document belong to class$\rightsquigarrow$ threshold to classify

Selecting λ

How do we determine λ? $\rightsquigarrow$ Cross validation (Recall code from Monday)
Applying models gives score (probability) of document belong to class$\rightsquigarrow$ threshold to classify

Credit Claiming (Grimmer, Westwood, and Messing 2014)

```
library(glmnet)
set.seed(8675309) ##setting seed
folds<- sample(1:10, nrow(dtm), replace=T) ##assigning to fold
out_of_samp<- c() ##collecting the predictions
```

Credit Claiming (Grimmer, Westwood, and Messing 2014)

```
for(z in 1:10){
train<- which(folds!=z) ##the observations we will use to train the model

test<- which(folds==z) ##the observations we will use to test the model
part1<- cv.glmnet(x = dtm[train,], y = credit[train], alpha = 1, family =
b́inomial)́ ##fitting the LASSO model on the data.
## alpha = 1 -> LASSO
## alpha = 0 -> RIDGE
## 0<alpha<1 -> Elastic-Net
out_of_samp[test]<- predict(part1, newx= dtm[test,], s = part1$lambda.min,
type =ćlass)́ ##predicting the labels
print(z) ##printing the labels
}
conf_table<- table(out_of_samp, credit) ##calculating the confusion table
> round(sum(diag(conf_table))/len(credit), 3)
[1] 0.844
```

Generalized Cross Validation and Ridge Regression

In some special cases there are analytic solutions:

Generalized Cross Validation and Ridge Regression

In some special cases there are analytic solutions:

$$\boldsymbol{\beta}^{\textsf{Ridge}} = \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J\right)^{-1} \boldsymbol{X}^{'}\boldsymbol{Y}$$

Generalized Cross Validation and Ridge Regression

In some special cases there are analytic solutions:

$$\begin{aligned}
\boldsymbol{\beta}^{\textsf{Ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
\widehat{\boldsymbol{Y}} &= \boldsymbol{X}(\boldsymbol{\beta})^{\textsf{Ridge}}
\end{aligned}$$

Generalized Cross Validation and Ridge Regression

In some special cases there are analytic solutions:

$$\begin{aligned}
\boldsymbol{\beta}^{\textsf{Ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J\right)^{-1} \boldsymbol{X}^{'}\boldsymbol{Y} \\
\widehat{\boldsymbol{Y}} &= \boldsymbol{X}(\boldsymbol{\beta})^{\textsf{Ridge}} \\
&= \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J\right)^{-1} \boldsymbol{X}^{'}}_{\textsf{Hat Matrix}}\boldsymbol{Y}
\end{aligned}$$

Generalized Cross Validation and Ridge Regression

In some special cases there are analytic solutions:

$$\begin{aligned}
\boldsymbol{\beta}^{\textsf{Ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
\widehat{\boldsymbol{Y}} &= \boldsymbol{X}(\boldsymbol{\beta})^{\textsf{Ridge}} \\
&= \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}}_{\textsf{Hat Matrix}}\boldsymbol{Y} \\
\widehat{\boldsymbol{Y}} &= \underbrace{\boldsymbol{H}}_{\textsf{Smoother Matrix}} \quad \boldsymbol{Y}
\end{aligned}$$

Generalized Cross Validation and Ridge Regression

In some special cases there are analytic solutions:

$$
\begin{aligned}
\boldsymbol{\beta}^{\textsf{Ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J\right)^{-1} \boldsymbol{X}^{'}\boldsymbol{Y} \\
\widehat{\boldsymbol{Y}} &= \boldsymbol{X}(\boldsymbol{\beta})^{\textsf{Ridge}} \\
&= \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J\right)^{-1} \boldsymbol{X}^{'}}_{\textsf{Hat Matrix}}\boldsymbol{Y} \\
\widehat{\boldsymbol{Y}} &= \underbrace{\boldsymbol{H}}_{\textsf{Smoother Matrix}} \quad \boldsymbol{Y}
\end{aligned}
$$

Generalized Cross Validation and Ridge Regression

Why do we care?

Generalized Cross Validation and Ridge Regression

Why do we care?
Leave one out cross validation

Generalized Cross Validation and Ridge Regression

Why do we care?
Leave one out cross validation

$$\text{Cross Validation}(1) \quad = \quad \frac{1}{N}\sum_{i=1}^{N}(Y_i - f(\boldsymbol{X}_{-i}, \boldsymbol{Y}_{-i}, \lambda, \hat{\boldsymbol{\beta}}))^2$$

Generalized Cross Validation and Ridge Regression

Why do we care?
Leave one out cross validation

$$\begin{aligned}
\text{Cross Validation}(1) &= \frac{1}{N}\sum_{i=1}^{N}(Y_i - f(\boldsymbol{X}_{-i}, \boldsymbol{Y}_{-i}, \lambda, \hat{\boldsymbol{\beta}}))^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - f(\boldsymbol{X}, \boldsymbol{Y}, \lambda, \hat{\boldsymbol{\beta}})}{1 - H_{ii}}\right)^2
\end{aligned}$$

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace$(\boldsymbol{H}) \equiv \text{Tr}(\boldsymbol{H}) = \sum_{i=1}^{N} H_{ii}$

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace$(\boldsymbol{H}) \equiv$ Tr$(\boldsymbol{H}) = \sum_{i=1}^{N} H_{ii}$
- Tr$(\boldsymbol{H})$ = Effective number of parameters (class regression = number of independent variables + 1)

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace$(\boldsymbol{H}) \equiv$ Tr$(\boldsymbol{H}) = \sum_{i=1}^{N} H_{ii}$
- Tr$(\boldsymbol{H}) =$ Effective number of parameters (class regression $=$ number of independent variables $+ 1$)
- For Ridge regression:

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace($\boldsymbol{H}$) $\equiv$ Tr($\boldsymbol{H}$) $= \sum_{i=1}^{N} H_{ii}$
- Tr($\boldsymbol{H}$) $=$ Effective number of parameters (class regression $=$ number of independent variables $+ 1$)
- For Ridge regression:

$$\text{Tr}(\boldsymbol{H}) \;=\; \sum_{j=1}^{J} \frac{\lambda_j}{\lambda_j + \underbrace{\lambda}_{\text{Penalty}}}$$

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace$(\boldsymbol{H}) \equiv$ Tr$(\boldsymbol{H}) = \sum_{i=1}^{N} H_{ii}$
- Tr$(\boldsymbol{H})$ = Effective number of parameters (class regression = number of independent variables + 1)
- For Ridge regression:

$$\text{Tr}(\boldsymbol{H}) \quad = \quad \sum_{j=1}^{J} \frac{\lambda_j}{\lambda_j + \underbrace{\lambda}_{\text{Penalty}}}$$

where λ_j is the j^{th} Eigenvalue from $\boldsymbol{\Sigma} = \boldsymbol{X}^{'}\boldsymbol{X}$

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace$(\boldsymbol{H}) \equiv$ Tr$(\boldsymbol{H}) = \sum_{i=1}^{N} H_{ii}$
- Tr$(\boldsymbol{H})$ = Effective number of parameters (class regression = number of independent variables + 1)
- For Ridge regression:

$$\text{Tr}(\boldsymbol{H}) \quad = \quad \sum_{j=1}^{J} \frac{\lambda_j}{\lambda_j + \underbrace{\lambda}_{\text{Penalty}}}$$

where λ_j is the j^{th} Eigenvalue from $\boldsymbol{\Sigma} = \boldsymbol{X}^{'}\boldsymbol{X}$ (!!!!!)

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace$(\boldsymbol{H}) \equiv$ Tr$(\boldsymbol{H}) = \sum_{i=1}^{N} H_{ii}$
- Tr$(\boldsymbol{H})$ = Effective number of parameters (class regression = number of independent variables + 1)
- For Ridge regression:

$$\text{Tr}(\boldsymbol{H}) \;=\; \sum_{j=1}^{J} \frac{\lambda_j}{\lambda_j + \underbrace{\lambda}_{\text{Penalty}}}$$

where λ_j is the j^{th} Eigenvalue from $\boldsymbol{\Sigma} = \boldsymbol{X}^{'}\boldsymbol{X}$ (!!!!!)

Define generalized cross validation:

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace$(\boldsymbol{H}) \equiv$ Tr$(\boldsymbol{H}) = \sum_{i=1}^{N} H_{ii}$
- Tr$(\boldsymbol{H})$ = Effective number of parameters (class regression = number of independent variables + 1)
- For Ridge regression:

$$\text{Tr}(\boldsymbol{H}) \quad = \quad \sum_{j=1}^{J} \frac{\lambda_j}{\lambda_j + \underbrace{\lambda}_{\text{Penalty}}}$$

where λ_j is the j^{th} Eigenvalue from $\boldsymbol{\Sigma} = \boldsymbol{X}^{'}\boldsymbol{X}$ (!!!!!)

Define generalized cross validation:

$$\text{GCV} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - \hat{Y}_i}{1 - \frac{\text{Tr}(\boldsymbol{H})}{N}}\right)^2$$

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace$(\boldsymbol{H}) \equiv$ Tr$(\boldsymbol{H}) = \sum_{i=1}^{N} H_{ii}$
- Tr$(\boldsymbol{H})$ = Effective number of parameters (class regression = number of independent variables + 1)
- For Ridge regression:

$$\text{Tr}(\boldsymbol{H}) \quad = \quad \sum_{j=1}^{J} \frac{\lambda_j}{\lambda_j + \underbrace{\lambda}_{\text{Penalty}}}$$

where λ_j is the j^{th} Eigenvalue from $\boldsymbol{\Sigma} = \boldsymbol{X}^{'}\boldsymbol{X}$ (!!!!!)

Define generalized cross validation:

$$\text{GCV} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - \hat{Y}_i}{1 - \frac{\text{Tr}(\boldsymbol{H})}{N}}\right)^2$$

Applicable in any setting where we can write Smoother matrix

Generalized Cross Validation and Ridge Regression

Calculating $\boldsymbol{H}$ can be computationally expensive

- Trace$(\boldsymbol{H}) \equiv$ Tr$(\boldsymbol{H}) = \sum_{i=1}^{N} H_{ii}$
- Tr$(\boldsymbol{H})$ = Effective number of parameters (class regression = number of independent variables + 1)
- For Ridge regression:

$$\text{Tr}(\boldsymbol{H}) = \sum_{j=1}^{J} \frac{\lambda_j}{\lambda_j + \underbrace{\lambda}_{\text{Penalty}}}$$

where λ_j is the j^{th} Eigenvalue from $\boldsymbol{\Sigma} = \boldsymbol{X}'\boldsymbol{X}$ (!!!!!)

Define generalized cross validation:

$$\text{GCV} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - \hat{Y}_i}{1 - \frac{\text{Tr}(\boldsymbol{H})}{N}}\right)^2$$

Applicable in any setting where we can write Smoother matrix