UNIVERSITY *of* DELAWARE

Modern Machine Learning
Ridge Regression

Kenneth E. Barner
Department of Electrical and Computer Engineering
University of Delaware

Ridge Regression

Methodology: Shrinkage Methods aim to reduce the variance and over fitting of least-squares methods by penalizing solutions with large numbers of nonzero coefficients

Ridge Regression

- Penalizes the coefficients in $\boldsymbol{\beta}$ to be smaller
- Formulation

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{(p+1)}}{\operatorname{argmin}} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \qquad (*)$$

Advantages:

- The regularization parameter λ can be used to avoid overfitting
- The solution can be written in a closed form

Note: Ridge Regression can be equivalently formulated as

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{(p+1)}}{\operatorname{argmin}} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 \qquad (**)$$

$$\text{subject to } \|\boldsymbol{\beta}\|^2 < t$$

There is a one-to-one correspondence between λ in $(*)$ and t in $(**)$

Ridge Regression

The residual sum of squares is denoted as

$$RSS(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

$$\begin{aligned} \frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + 2\lambda\mathbf{I}\boldsymbol{\beta} \\ &= 2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} - 2\mathbf{X}^T\mathbf{y} = \mathbf{0} \end{aligned}$$

Therefore

$$\widehat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- When $\lambda = 0$, we have the least squares solution

Ridge Regression

Singular value decomposition (SVD) analysis

Assume that the data, $\mathbf{X} \in \mathbb{R}^{N\times p}$, has been centered, $x_{ij} \leftarrow (x_{ij} - \overline{x_j})$

Then we estimate β_0 by $\bar{y} = \frac{1}{N}\sum_{i=1}^{p} y_i$, and perform ridge regression without intercept ($\mathbf{X} \in \mathbb{R}^{N\times p}$)

According to SVD theory we have that $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where:

$\mathbf{U} \in \mathbb{R}^{N\times p}$ and $\mathbf{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrices

$\mathbf{D} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with diagonal entries

$d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$, which are called singular values of $\mathbf{X}$

Ridge Regression

Analysis: compare the LS and Ridge solutions

$$\widehat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\mathbf{X}^T\mathbf{y} \quad ; \quad \widehat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Using $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ we obtain

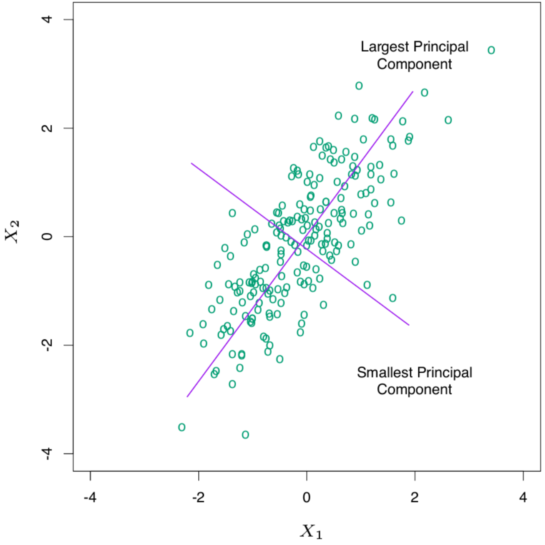
$$\widehat{\boldsymbol{\beta}}_{LS} = \sum_{i=1}^{p} \mathbf{u}_j \underline{\mathbf{u}_j^T\mathbf{y}} \quad ; \quad \widehat{\boldsymbol{\beta}}_{ridge} = \sum_{i=1}^{p} \mathbf{u}_j \left(\frac{d_j^2}{d_j^2 + \lambda}\right) \underline{\mathbf{u}_j^T\mathbf{y}}$$

Observations:

- Since $\lambda \geq 0$, the factor $\left(\frac{d_j^2}{d_j^2+\lambda}\right) \leq 1$
- Ridge regression computes the coordinates of $\mathbf{y}$ in the orthonormal basis $\mathbf{U}$, and then shrinks the components by $d_j^2/(d_j^2 + \lambda)$
- For vectors with low d_j^2, the shrinkage in ridge regression is higher
- SVD is directly related to principal component decomposition $\Longrightarrow$ the largest shrinkage is applied to coefficients along the eigenvector direction with the lowest eigenvalue

Ridge Regression

- The vectors associated with high values d_j represent the directions where the data has the biggest variance (also called principal components)
- Therefore, ridge regression reduces the effect of noisy data in the directions other than those of the principal components



Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects y onto these components, and then shrinks the coefficients of the low variance components more than the high-variance components.

Example: Prostate Cancer

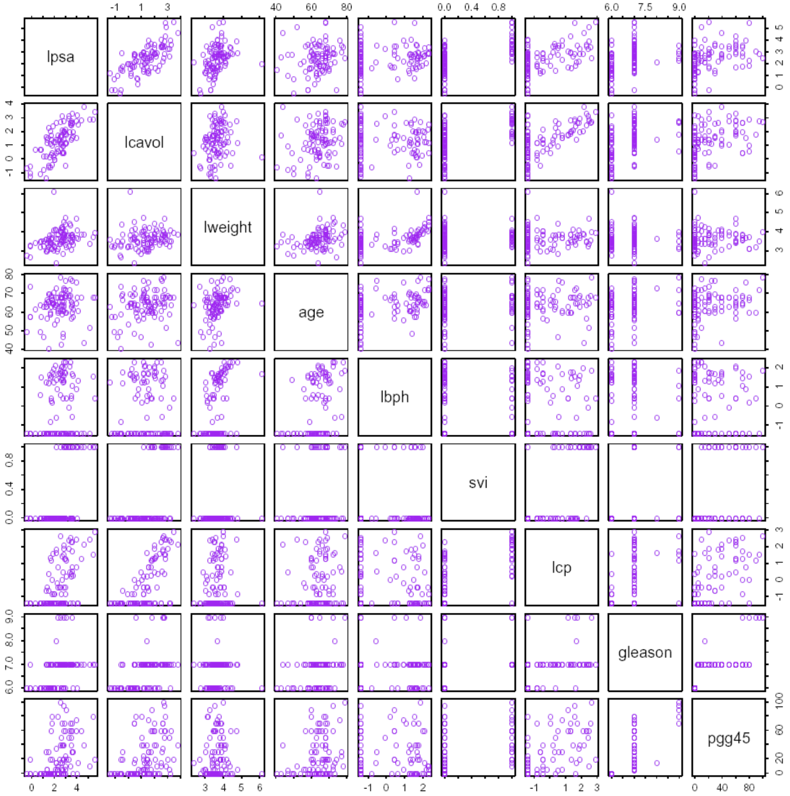
Example: Stanley et al. (1989) examined the correlation between prostate-specific antigen and multiple clinical measures in prostatectomy patients.

Data: Given variables (clinical measures):

`lcavol`: log cancer volume
`lweight`: log prostate weight
`age`
`lbph`: log benign hyperplasia amount
`svi`: seminal vesicle invasion
`lcp`: log capsular penetration
`gleason`: gleason score
`pgg45`: percent gleason scores 4 or 5

Objective: Predict `lpsa` (log of prostate specific antigen) level

Data size: Measurements from 97 men



Pairwise scatterplot matrix of data. First row is `lpsa` versus each clinical measure. Note `svi` and `gleason` are categorical.

Example: Prostate Cancer (continued)

- Fit a ridge regression model to `lpsa`
- Training/testing split: 67/30
- Plot the ridge as functions of $df(\lambda)$, the effective degrees of freedom implied by the penalty λ

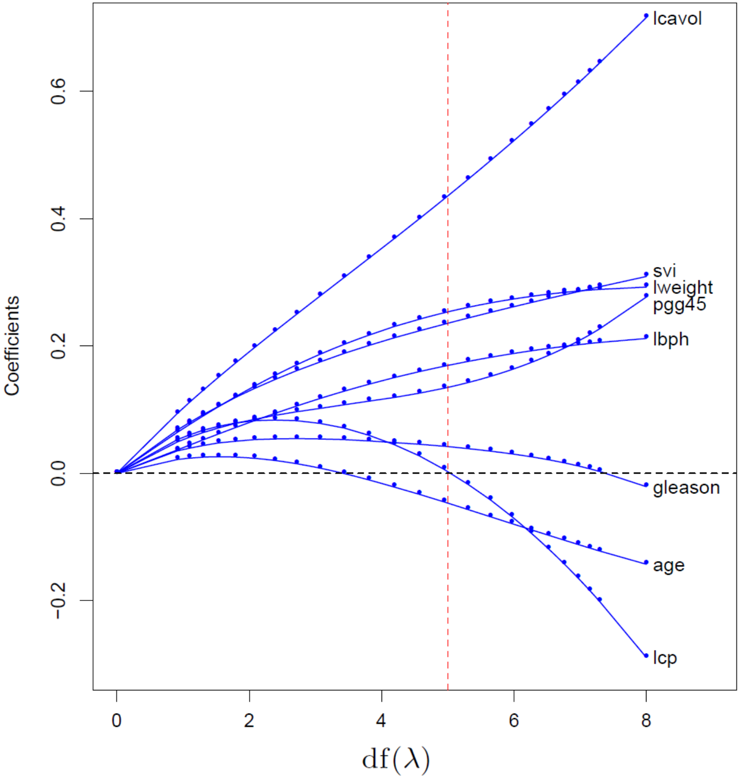
$$df(\lambda) = tr[\boldsymbol{X}(\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{I})^{-\mathbf{1}})\mathbf{X}^T] = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

- Note:

  $df(\lambda)$= p when λ = 0 (no regularization)

  $df(\lambda) \rightarrow 0$ as $\lambda \rightarrow \infty$.

- The minimum error, as determined through cross validation, occurs at $df(\lambda)$ = *5.0*.



Profiles of ridge coefficients for the prostate cancer example