Data Science II
(P8106)

Department of Biostatistics
Mailman School of Public Health
Columbia University

Spring 2024

Shrinkage methods / regularization

- Recall
  - Ordinary least squares: bias? variance?
- Why not least squares?
  - Too many predictors, e.g., $p > n$
  - Collinearity $\rightarrow$ large variance
- Bias-variance trade-off
  - A slight increase in bias but lower variance
- To control variance, we may **regularize** the coefficient, i.e., control how large the coefficients grow
- Ridge regression, the lasso, and the elastic net

Ridge regression

- Recall that the least squares estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize
  $RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$
- The ridge regression coefficient estimates $\hat{\beta}_\lambda^R$ are the values that minimize

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

  where $\lambda \geq 0$ is a **tuning parameter**
  - $\lambda = 0$?
  - $\lambda = \infty$?

Ridge regression

- The second term, $\lambda \sum_{j=1}^{p} \beta_j^2$ is called a shrinkage penalty
  - The penalty is small when $\beta_1, \ldots, \beta_p$ are close to zero
  - λ control the relative impact of these two terms on the regression coefficient estimates
  - The intercept is unpenalized
- Selecting a good value of λ is critical
- Balancing two ideas: fitting a linear model and shrinking the coefficients

Selecting the tuning parameter λ

- Cross-validation
  - Choose a grid of λ values
  - Compute the cross-validation error rate for each value of λ
  - Select the tuning parameter value for which the cross-validation error is smallest
- The model is re-fit using all of the available observations and the selected value of λ

Standardizing predictors

- Least squares: Multiplying X_j by a constant c simply leads to a scaling of coefficient estimates by a factor of $1/c$
- The ridge regression coefficient estimates can change substantially when multiplying a predictor by c
- The penalty is unfair if the predictor variables are not on the same scale
- Apply ridge regression after *scaling* the predictors (to have sample variance 1)
- If we *center* the predictors (to have sample mean 0), the intercept estimate ends up being $\hat{\beta}_0 = \bar{y}$
- One can center $y, X \rightarrow$ no intercept

Why does ridge regression improve over least squares?

- As λ increases, bias? variance?
- Is ridge regression helpful when all the true coefficients are large?
  - Ridge regression performs particularly well when there is a subset of true coefficients that are small
  - When all of the true coefficients are moderately large, it can still outperform linear regression over a pretty narrow range of small λ values

Why does ridge regression improve over least squares?

- As λ increases, bias? variance?
- Is ridge regression helpful when all the true coefficients are large?
  - Ridge regression performs particularly well when there is a subset of true coefficients that are small
  - When all of the true coefficients are moderately large, it can still outperform linear regression over a pretty narrow range of small λ values

The ridge solution $\hat{\beta}_\lambda^R$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

- The design matrix $\boldsymbol{X}$ is assumed to be standardized
- The response vector $\boldsymbol{y}$ is assumed to be centered

$$\hat{\beta}_\lambda^R = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Understanding ridge from singular value decomposition

$$
\begin{aligned}
\boldsymbol{X} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\intercal \\
\hat{\beta}_\lambda^R &= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{D}\boldsymbol{U}^\intercal \boldsymbol{y} \\
\hat{y}_\lambda^R &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{D}\boldsymbol{U}^\intercal \boldsymbol{y}
\end{aligned}
$$

Degree of freedom for ridge regression

- A smoother matrix $\boldsymbol{S}$ is a linear operator satisfying $\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$
- Effective degree of freedom is $tr(\boldsymbol{S}) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}$

Lasso

- How does ridge regression perform if a group of the true coefficients was exactly zero?
- Ridge regression will include all p predictors in the final model
- Ridge regression does poorly in terms of offering a clear interpretation
- The least absolute shrinkage and selection operator (lasso) is an alternative to ridge regression that overcomes this disadvantage

Lasso

- The lasso coefficients $\hat{\beta}_\lambda^L$, minimizes the quantity

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- The lasso uses a ℓ_1 penalty instead of a ℓ_2 penalty
- Standardize the predictors before fitting the lasso. Why?

Lasso

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero
- The ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be zero when λ is sufficiently large
- There is no analytic solution for the lasso
- The solution is nonlinear in $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$
- The lasso performs variable selection and yields sparse models (i.e., models that involve only a subset of the variables)
- Selecting a good value of λ for the lasso is critical - cross validation

The lasso and ridge regression

One can show that

- The lasso coefficient estimates solve

$$\text{minimize}_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

- The ridge regression coefficient estimates solve

$$\text{minimize}_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s$$

The lasso and ridge regression

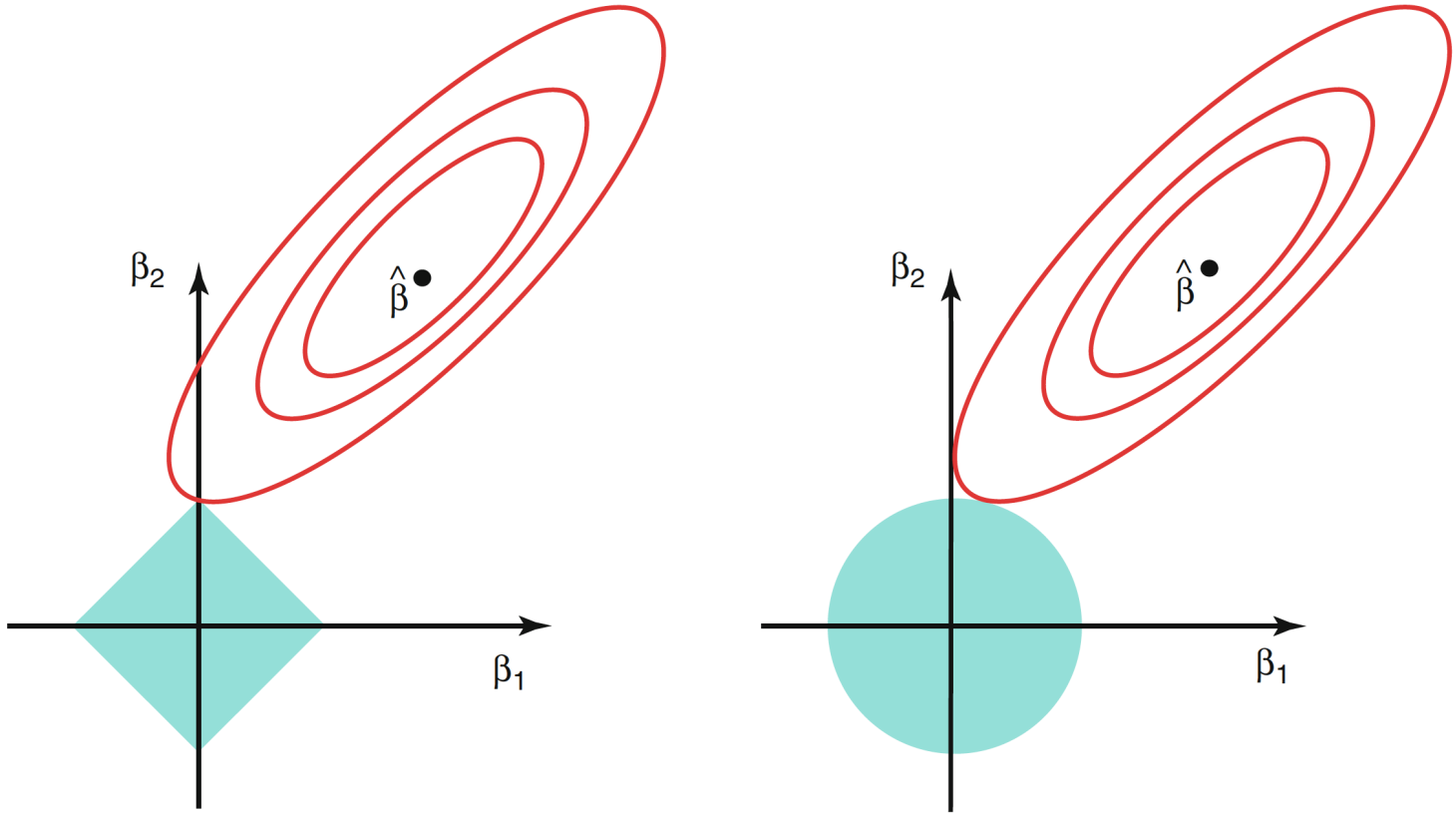

Figure: ISL 6.7

The lasso performs ℓ_1 shrinkage, so that there are "corners" in the constraint. If the sum of squares "hits" one of these corners, the coefficient corresponding to the axis is shrunk to zero.

Comparing the two types of penalties

- Ridge regression is known to shrink the coefficients of correlated predictors towards each other, allowing them to borrow strength from each other
- In the extreme case of p identical predictors, they each get identical coefficients with $1/p$th the size that any single one would get if fit alone
- Lasso is somewhat indifferent to very correlated predictors, and will tend to pick one and ignore the rest

Summary

- Neither ridge regression nor the lasso will universally dominate the other
- When the response is a function of only a relatively small number of predictors, which to use?
- However, the number of predictors that is related to the response is never known for real data sets
- How to determine which approach is better on a particular data set?

Elastic net

- Minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{i=1}^{n}\beta_j x_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{p}\beta_j^2 + \lambda_2 \sum_{j=1}^{p}|\beta_j|$$

- Effective regularization via the ridge-type penalty
- Feature selection via the lasso penalty
- More effective to deal with groups of highly correlated predictors