The Lasso

DS 301

Iowa State University

See R script: `shrinkage_methods.R`

Why does ridge regression improve over least squares?

Ridge regression's advantage over least square is rooted in the **bias-variance trade-off**.

- As λ increases, the flexibility of the ridge regression fit decreases, leading to a decreased variance but increased bias.

Ridge regression recap

- Minimizes the usual regression criterion (RSS) plus a l_2 penalty term.
- It can shrink coefficients towards 0 by introducing some bias.
- This can improve prediction.
- (Works well in the presence of multicollinearity.)
- Amount of shrinkage is controlled by λ. $\longrightarrow$ λ using CV.
- Ridge regression performs particularly well when there is a subset of true regression coefficients that are **small** or even **zero.**

Disadvantage of ridge regression

can never set regression coefficients to be exactly 0.

↳ will always return to you the full model.

The Lasso (least absolute selection & shrinkage operator)

- Resolve disadvantage of ridge regression

- Performs **both** model selection and regularization.

||
can set regression coefficients
to be exactly 0.

The Lasso

We want regression coefficients $\hat{B}_{lasso}$ such that

$$\hat{B}_{lasso} = \min_{B}\left(\sum_{i=1}^{n} \left(y_i - (B_0 + B_1 X_1 + \cdots B_p X_p)\right)^2 + \underbrace{\lambda \sum_{j=1}^{p} |B_j|}_{l_1 \text{ penalty}} \right)$$

$\lambda = 0 \Rightarrow \hat{B}_{lasso}$ defaults to least squares

$\lambda = \infty \Rightarrow \hat{B}_{lasso} = 0.$

The Lasso

For a λ in between the extreme, we are balancing two ideas:

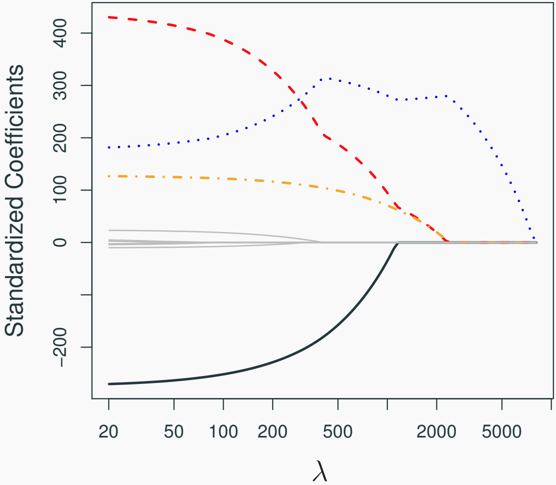
- Fitting a linear model of Y on X.
- Shrinking the coefficients (l_1 penalty can shrink some to 0).

Lasso has no analytical solution (no closed-form formula).

- Can find lasso regression coefficients using numerical algorithms.

  (gradient descent,
  newton raphson, etc.)

Lasso regression coefficients



Lasso vs. Ridge

- In terms of prediction error (test MSE), the lasso performs comparably to ridge regression.
- Lasso penalty can set some coefficients to 0 when λ is sufficiently large.
  - Performs automatic model selection.
  - Leads to sparse models. (less predictors)
- Selecting λ here is (again) critical and can be done using cross-validation.
- Lasso implicitly assumes that a number of the coefficients truly equal zero.

Lasso vs. Ridge

- Neither ridge regression nor the lasso will universally dominate the other.
- Lasso will generally perform better when a relatively small number of predictors have substantial coefficients.
- Ridge regression will generally perform better when the response is a function of many predictors.
- The number of predictors that is related to the response is almost **never known** beforehand for real data sets.
- Cross-validation can help us determine which approach is better on a particular data set.

See R script: `shrinkage_method.R`

Predictive modeling tools at your disposal

(1) Least squares linear regression lm(·)

- Analytical solution, simple to implement.
- Model non-linear relationships (polynomial regression, regression splines, natural splines)
- Incorporate higher order terms (interactions).
- Model selection (subset selection, forward, backward, stepwise, cross-validation) regsubsets(·)
- Inference is straightforward to carry out. $\longrightarrow$ • multicollinearity

However, unbiased estimates of $\hat{B}$: $E(\hat{B}) = B$.

↳ potentially high variance

$\Longrightarrow$ we don't get best prediction error

* $\uparrow$ bias $\rightarrow$ $\downarrow$ variance $\rightarrow$ $\downarrow$ best MSE.

Predictive modeling tools at your disposal

(2) Ridge Regression

- Useful in improving prediction accuracy.
- Will always result in a full model with all p predictors. Ridge regression is the obvious choice if you believe all predictors are somewhat important.
- Can handle multicollinearity.
- Inference can also be done (relatively straightforwardly).

(3) The Lasso

- Regularizes and performs model selection.
- Generally works well when only a subset of predictors are actually important.
- Inference not as straightforward to carry out.

No one approach will universally dominate the other.

Extensions of Lasso

- elastic net
    - addresses some of shortcomings of lasso
        - (does not do well in presence of multicollinearity)
        - does not work well when $p \geq n$. (high dimensional)

↳ takes best of both worlds:

$$\min_{B} \left(\sum_{i=1}^{n} (y_i - (B_0 + B_1 x_1 + \cdots B_p x_p))^2 + \underbrace{\lambda_1 \sum_{j=1}^{p} |B_j|}_{\substack{\text{lasso penalty} \\ (l_1)}} + \underbrace{\lambda_2 \sum_{j=1}^{p} B_j^2}_{\substack{\text{ridge penalty.} \\ (l_2)}} \right).$$

Extensions of Lasso

- group lasso
  - ↳ categorical predictors (K) → K-1 dommy variables.

→ allows groups of predictors to be selected in/out of model together.

⇒ keep collection of dommy variables together.

categorical predictors ✓

biological studies ✓