Automatic Model Selection

Roberta De Vito

BROWN
Public Health

Shrinkage Methods

- Subset selection methods $\rightarrow$ least squares to fit a model with subset of the predictors

Shrinkage Methods

- Subset selection methods $\rightarrow$ least squares to fit a model with subset of the predictors
- Alternatively $\rightarrow$ techniques that constrain or regularize coefficient estimates
- Shrinking coefficient estimates can significantly reduce their variance

Ridge Regression

- Ordinary Least Squares (OLS) estimates β's by minimizing RSS
- Ridge regression minimizes RSS penalized by the size of the regression coefficients,

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

λ is a *tuning parameter* controlling the size of the penalty (usually selected using cross-validation).

Question on Prismia

Shrinkage

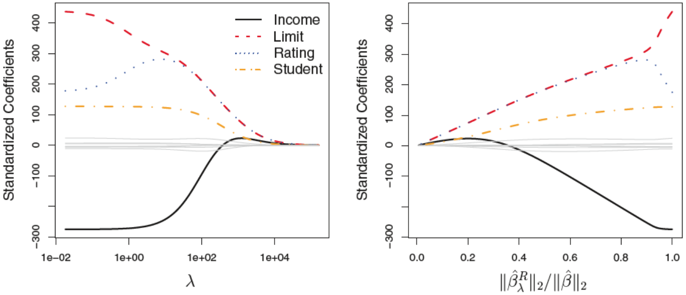
1. Fit data with the RSS
2. Tuning parameter λ control amount of shrinkage
3. $\lambda \sum_{j=1}^{p} \beta_j^2 \rightarrow$ shrinkage penalty
4. Shrinkage does not involve intercept
5. No shrinkage in the intercept
6. Standardize the covariates $\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{1/n(x_{ij} - \bar{x}_j)^2}}$

Shrinkage

- Selecting a good value for λ is critical: cross-validation
- Because penalty term involves sizes of coefficients, it depends on scale of variables
- Need to standardize variables by dividing each by their standard deviation

The Credit Data

	Coefficient	Std.error	p-value
Intercept	-173.4	43.8	< 0.0001
Income	-299.2	54.2	< 0.0001
Limit	432.3	75.6	< 0.0001
Rating	176.3	42.6	< 0.0001
Student	112.2	33.2	< 0.0001



Bias Variance Trade-off I

As λ increases do you think that the variance of the ridge regression fit will

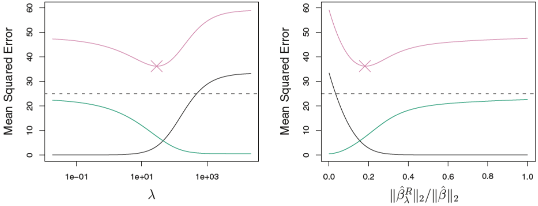
- increase
- decrease

Bias Variance Trade-off I

As λ increases do you think that the bias of the ridge regression fit will

- increase
- decrease

Bias Variance Trade-off II



Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set.

Why Does Ridge Regression Improve Over Least Squares?

Why Does Ridge Regression Improve Over Least Squares?

- Works when $p > n$.
- Reduces variance by introducing bias into the model

Lasso

- Penalty term for ridge regression: coefficients not exactly 0
- Lasso: force coefficients to be zero

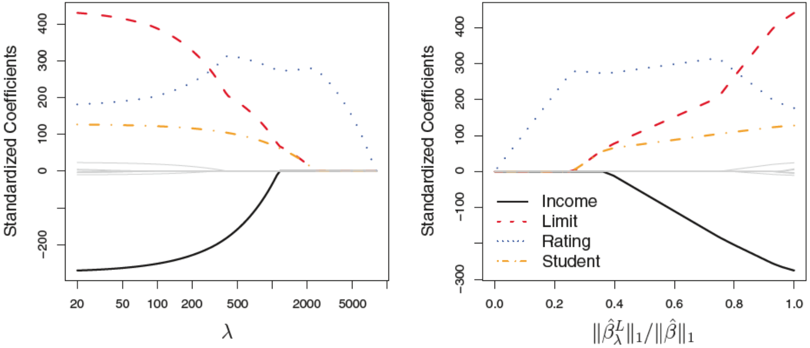
$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Penalty is ℓ_1 rather than ℓ_2

Lasso

- Lasso shrinks coefficient estimates towards zero
- ℓ_1 penalty force some of coefficient estimates to be exactly equal to zero
- Sparsity and interpretation
- As in ridge regression, selecting a good value of λ for lasso is critical: cross-validation

Example of Lasso Shrinkage



Another Formulation of Minimization for Ridge and Lasso

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} s.t. \sum_{j=1}^{p} \beta_j^2 \leq s$$

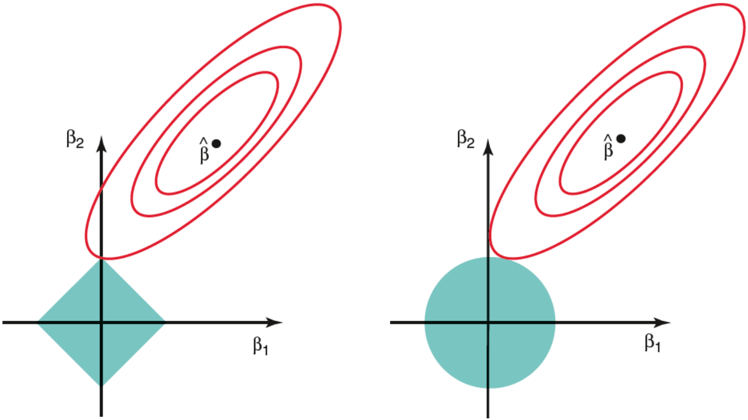
and

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} s.t. \sum_{j=1}^{p} |\beta_j| \leq s$$

- s and λ are constraints
- s imposes budget on how large coefficients can be

Variable Selection Property of Lasso

Contours of RSS (red)
Constraint regions (blue) for lasso (left) and ridge (right)

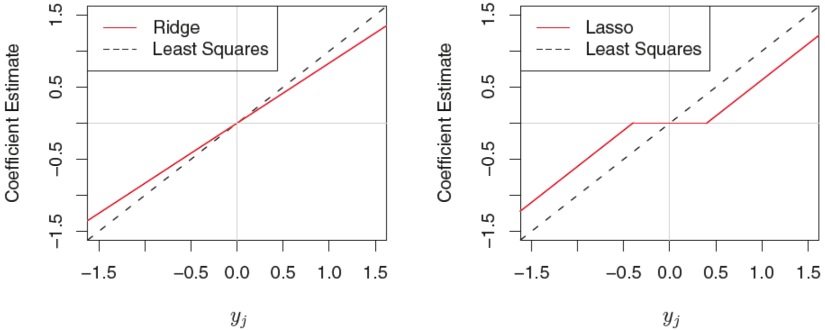


$p = 3 \rightarrow$ constraint region for ridge regression is a sphere, constraint region for the lasso is a polyhedron

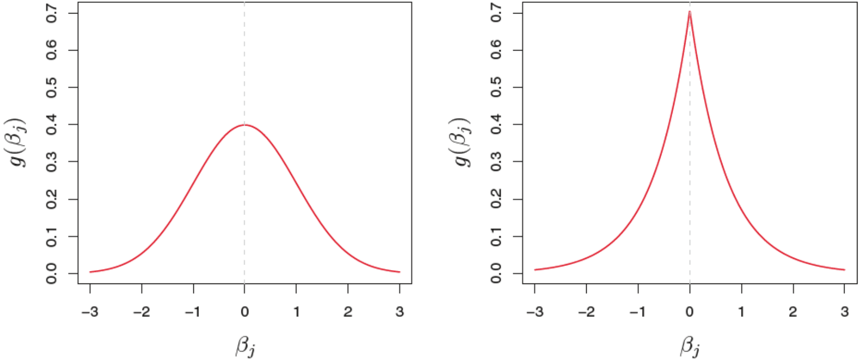
Choosing Shrinkage Method

- In general, lasso will tend to perform better when response is a function of only a relatively small number of predictors
- But number of predictors related to response is never known a priori for real data sets
- A technique such as cross-validation can be used to determine which approach is better on a particular data set

A better Intuition



Bayesian Interpretation for Ridge Regression and the Lasso



Left: Ridge regression as a Gaussian prior. Right: The lasso as a double-exponential prior.

Selecting Tuning Parameter by Cross-Validation

- Ridge regression and Lasso: methods to determine which model is best
- Method selecting value for tuning parameter λ
- How can we do that?

Selecting Tuning Parameter by Cross-Validation

- Ridge regression and Lasso: methods to determine which model is best
- Method selecting value for tuning parameter λ
- How can we do that?
- Cross-validation $\rightarrow$ grid of λ values, and compute cross-validation error rate for each value of λ
- Select λ for which cross-validation error is smallest
- Re-fit model using all available observations

Example: The Hitters data set

`AtBat`	Number of times at bat in 1986
`Hits`	Number of hits in 1986
`HmRun`	Number of home runs in 1986
`Runs`	Number of runs in 1986
`RBI`	Number of runs batted in in 1986
`Walks`	Number of walks in 1986
`Years`	Number of years in the major leagues
`CatBat`	Number of times at bat during his career
`CHits`	Number of hits during his career
`CHmRun`	Number of home runs during his career
`CRuns`	Number of runs during his career
`CRBI`	Number of runs batted in during his career
`CWalks`	Number of walks during his career
`League`	A and N player's league at the end of 1986
`Division`	E and W player's division at the end of 1986
`PutOuts`	Number of put outs in 1986
`Assists`	Number of assists in 1986
`Errors`	Number of errors in 1986
`Salary`	1987 annual salary in thousands of dollars
`NewLeague`	A and N levels player's league at the beginning of 1987

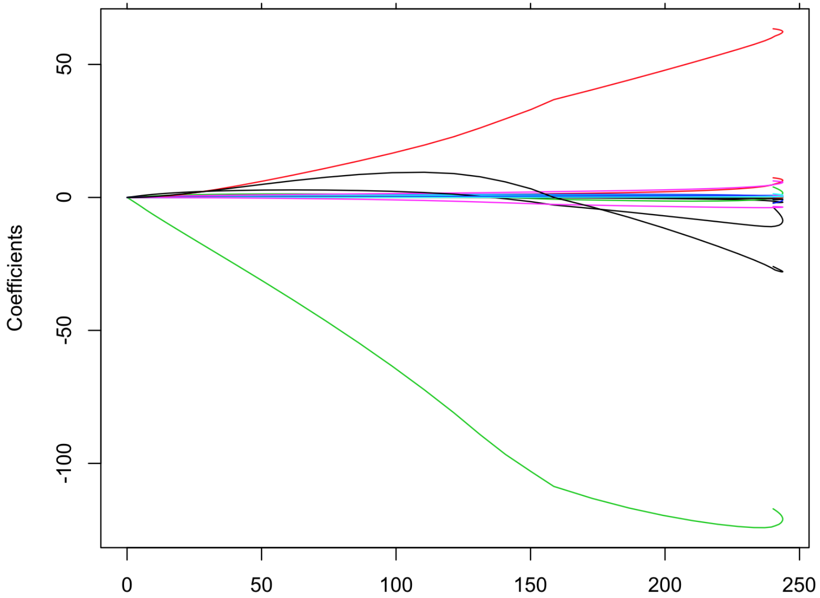
First look to the data

```
> head(Hitters)
                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI
-Alan Ashby          315   81     7   24  38    39    14   3449   835     69   321  414
-Alvin Davis         479  130    18   66  72    76     3   1624   457     63   224  266
-Andre Dawson        496  141    20   65  78    37    11   5628  1575    225   828  838
-Andres Galarraga    321   87    10   39  42    30     2    396   101     12    48   46
-Alfredo Griffin     594  169     4   74  51    35    11   4408  1133     19   501  336
-Al Newman           185   37     1   23   8    21     2    214    42      1    30    9
                   CWalks League Division PutOuts Assists Errors Salary NewLeague
-Alan Ashby           375      N        W     632      43     10  475.0         N
-Alvin Davis          263      A        W     880      82     14  480.0         A
-Andre Dawson         354      N        E     200      11      3  500.0         N
-Andres Galarraga      33      N        E     805      40      4   91.5         N
-Alfredo Griffin      194      A        W     282     421     25  750.0         A
-Al Newman             24      N        E      76     127      7   70.0         A
```

Linear Regression

```
Coefficients:
               Estimate
(Intercept)   163.10359
xAtBat         -1.97987
xHits           7.50077
xHmRun          4.33088
xRuns          -2.37621
xRBI           -1.04496
xWalks          6.23129
xYears         -3.48905
xCAtBat        -0.17134
xCHits          0.13399
xCHmRun        -0.17286
xCRuns          1.45430
xCRBI           0.80771
xCWalks        -0.81157
xLeagueN       62.59942
xDivisionW   -116.84925
xPutOuts        0.28189
xAssists        0.37107
xErrors        -3.36076
xNewLeagueN   -24.76233
```

Ridge Regression Coefficients



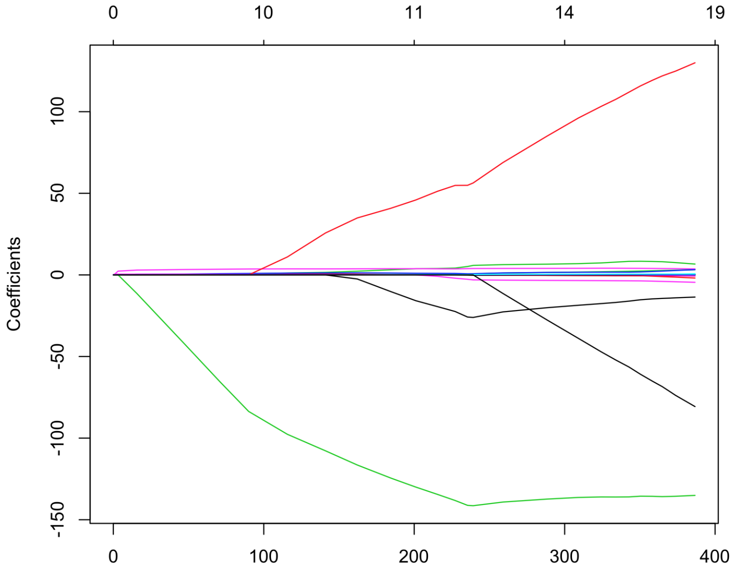
CV to choose the best lambda: Ridge

```
> cv.out = cv.glmnet(x_train, y_train, alpha = 0)
>
> ## Select lamda that minimizes training MSE
>  bestlam = cv.out$lambda.min
>
> bestlam
[1] 154.9432
```

The Ridge Regression

```
> predict(out, type = "coefficients", s = bestlam)[1:20,]
 (Intercept)        AtBat         Hits        HmRun         Runs          RBI 
 11.755707029 -0.017512197  1.143281649 -0.222569347  1.140681463  0.864182262 
        Walks        Years       CAtBat        CHits       CHmRun        CRuns 
  1.949383621 -0.996062806  0.010741238  0.071535200  0.488390858  0.141673578 
         CRBI       CWalks      LeagueN    DivisionW      PutOuts      Assists 
  0.150961984  0.004148626 31.056657299 -99.569019782  0.207911521  0.055049779 
       Errors   NewLeagueN 
 -2.154842352  4.759677438 
```

Lasso Coefficients



CV to choose the best lambda: Lasso

```
> cv.out = cv.glmnet(x_train, y_train, alpha = 1)
>
> # Select lambda that minimizes training MSE
> bestlam = cv.out$lambda.min
> bestlam
[1] 7.711561
```

The Lasso

```
> lasso_coef
(Intercept)        AtBat         Hits        HmRun         Runs          RBI
 9.74475651  -0.20002473   2.58693812   0.00000000   0.00000000   0.00000000
      Walks        Years       CAtBat        CHits       CHmRun        CRuns
 2.45127305  -1.32830861   0.00000000   0.00000000   0.06412011   0.24615220
       CRBI       CWalks      LeagueN    DivisionW      PutOuts      Assists
 0.41526531  -0.02681053  23.92281284 -117.94934641   0.24176933   0.00000000
     Errors   NewLeagueN
-0.94879450   0.00000000
```