Lecture 7: Model Selection and Regularization

STATS 202: Data Mining and Analysis

Linh Tran

tranlm@stanford.edu

Department of Statistics
Stanford University

July 14, 2021

Announcements

- HW2 due Friday.
- Midterm is on Wednesday.
  - Review this Friday
  - Open book (ISL/ESL)
  - No calculators necessary
  - Practice midterm solutions will be released tonight

- Subset selection
- Shrinkage methods
    - Ridge
    - LASSO
    - Elastic net

What we know so far

- In linear regression, adding predictors always decreases the training error or RSS.

$$RSS = (\mathbf{Y} - \mathbf{X}\beta)^\top(\mathbf{Y} - \mathbf{X}\beta) \tag{1}$$

- We can estimate β by minimizing the RSS.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y \tag{2}$$

- However, adding predictors does not necessarily improve the test error.
- Selecting significant predictors is hard when n is not much larger than p.

Singular matrices

- When our matrix is not of *full column rank* (e.g. $n < p$), we have that $(\mathbf{X}^\top \mathbf{X})^{-1}$ is not invertible.
- Consequently, there is no least squares solution:

$$\hat{\beta} = \underbrace{(\mathbf{X}^\top \mathbf{X})}_{\text{Singular}}{}^{-1} \mathbf{X}^\top y \tag{3}$$

  - So, we must find a way around this.

Accounting for singularity

$$\hat{\beta} = \underbrace{(\mathbf{X}^\top \mathbf{X})}_{\text{Singular}}{}^{-1}\mathbf{X}^\top y \tag{4}$$

Three common approaches for dealing with this:

1. Subset selection
   - Select a subset k of the p predictors ($k \leqslant p$).
   - Use criteria to help select which subset k we want.

Accounting for singularity

$$\hat{\beta} = \underbrace{(\mathbf{X}^\top \mathbf{X})}_{\text{Singular}}{}^{-1} \mathbf{X}^\top y \tag{4}$$

Three common approaches for dealing with this:

1. Subset selection
    - Select a subset k of the p predictors ($k \leqslant p$).
    - Use criteria to help select which subset k we want.
2. Shrinkage methods
    - Constrain the parameters we're estimating in some way

Accounting for singularity

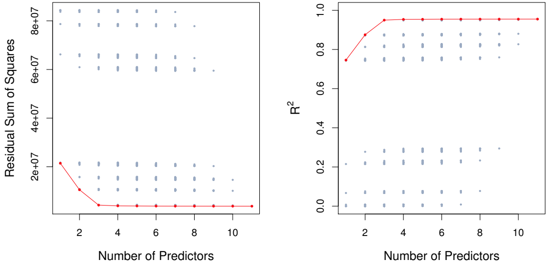
$$\hat{\beta} = \underbrace{(\mathbf{X}^\top \mathbf{X})}_{\text{Singular}}{}^{-1} \mathbf{X}^\top y \tag{4}$$

Three common approaches for dealing with this:

1. Subset selection
   - Select a subset k of the p predictors ($k \leqslant p$).
   - Use criteria to help select which subset k we want.
2. Shrinkage methods
   - Constrain the parameters we're estimating in some way
3. Dimension reduction
   - Project all our predictors to a smaller dimension space
   - Not covered in this class

Subset selection

- *Simple idea*: Compare all models with k predictors
- **Note**: There are $\binom{p}{k} = p!/(k!(p-k)!)$ possible models
- Choose the model with the smallest RSS
  - Doing this for every possible k:



Note: As expected, the RSS and R^2 improve with higher k.

The optimal k

Two approaches:

1. Use a hold out set (e.g. validation or test set)
   - c.f. Cross-validation
2. Use *modified* metrics that account for the size of k, e.g.
   - Akaike Information Criterion (AIC)
   - Bayesian Information Criterion (BIC)
   - Adjusted R^2

The optimal k

Two approaches:

1. Use a hold out set (e.g. validation or test set)
    - c.f. Cross-validation
2. Use *modified* metrics that account for the size of k, e.g.
    - Akaike Information Criterion (AIC)
    - Bayesian Information Criterion (BIC)
    - Adjusted R^2

How the modified metrics compare to using hold out sets

- Can be (much) less expensive to compute
- Motivated by asymptotic arguments and rely on model assumptions (e.g. normality of the errors)
- Equivalent concepts for other models (e.g. logistic regression)

Akaike Information Criterion (AIC)

Similar to Mallow's C_p:

$$C_p = \frac{1}{n}(RSS + 2k\hat{\sigma}^2) \tag{5}$$

- i.e. Adds the penalty $2k\hat{\sigma}^2$ to the RSS
- Can be shown to be unbiased estimate of test set error

But, also normalizes for $\hat{\sigma}^2$:

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2k\hat{\sigma}^2) = \frac{C_p}{\hat{\sigma}^2} \tag{6}$$

Since the two are proportional, (for least squares models) both are optimized at the same k.

Bayesian Information Criterion (BIC)

Similar to Mallow's C_p, but derived from Bayesian POV:

$$BIC = \frac{1}{n}(RSS + \log(n)k\hat{\sigma}^2) \tag{7}$$

n.b. $\log(n) > 2$ for $n > 7$

- BIC will penalize more for large k (i.e. optimizes for smaller k)

Adjusted R^2

Recall:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{8}$$

The adjusted R^2 penalizes for larger k:

$$R^2_{adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{9}$$

Adjusted R^2

Recall:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{8}$$

The adjusted R^2 penalizes for larger k:

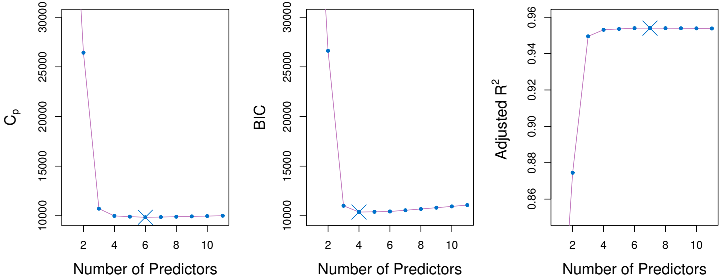
$$R^2_{adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{9}$$

Maximizing R^2_{adj} is equivalent to minimizing $1 - R^2_{adj}$, i.e.:

$$\frac{RSS}{n-d-1} \tag{10}$$

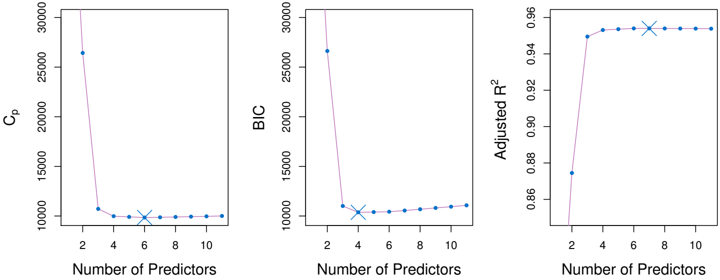
Applied example

Best subset selection for the Credit data set



Applied example

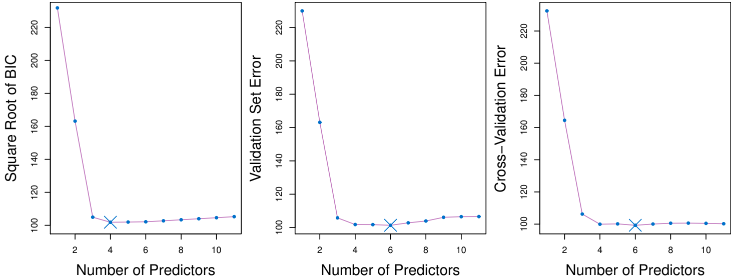
Best subset selection for the Credit data set



n.b. The curve is pretty flat for $k \geq 4$

Applied example

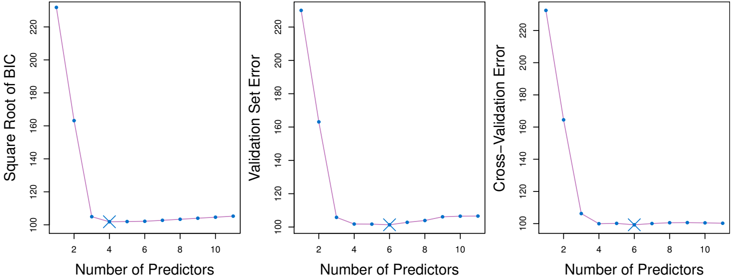
BIC vs validation sets



n.b. The curves are also pretty flat for $k \geq 4$.

Applied example

BIC vs validation sets



n.b. The curves are also pretty flat for $k \geq 4$.

Can use the *one-standard-error rule*

- Choose the parsimonious model (i.e. lowest k) such that the test error is within 1-SE of the lowest point

Stepwise selection methods

Best subset selection has 2 problems:

1. It is often very expensive computationally. We have to fit 2^p different models!
2. If for a fixed k, there are too many possibilities, we increase our chances of overfitting
   - i.e. the model selected has high variance.

Stepwise selection methods

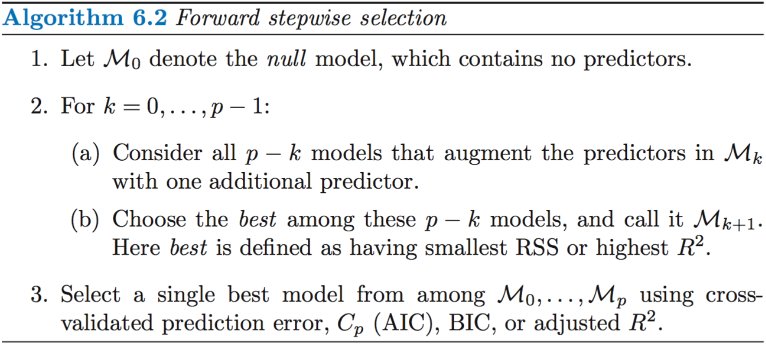
Best subset selection has 2 problems:

1. It is often very expensive computationally. We have to fit 2^p different models!
2. If for a fixed k, there are too many possibilities, we increase our chances of overfitting
   - i.e. the model selected has high variance.

One solution: Restrict our search space for the best model

- This reduces the variance of the selected model at the expense of an increase in bias.

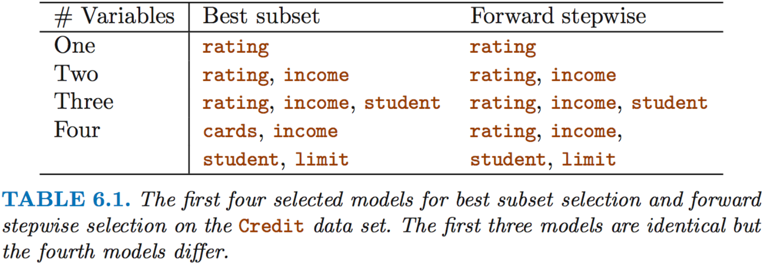
Forward stepwise selection



Algorithm 6.2 *Forward stepwise selection*

```
1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p - 1$:
    (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
    (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```

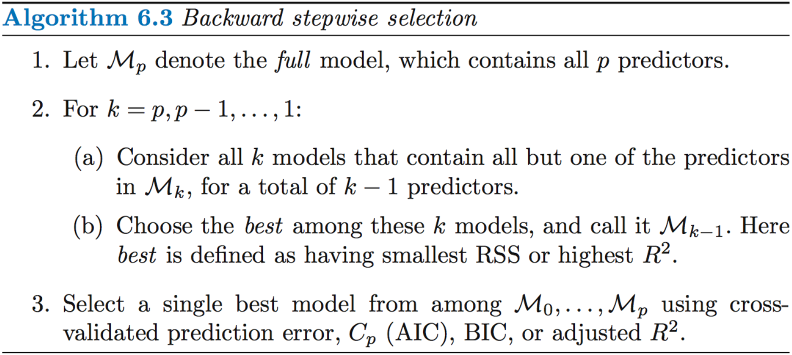
Forward selection vs best subset



# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income student, limit	rating, income, student, limit

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* Credit *data set. The first three models are identical but the fourth models differ.*

Backward stepwise selection



Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   - (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   - (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward vs backward selection

- You cannot apply backward selection when $p > n$
  - Though should still have a “reasonable” number of observations
- **Important**: they may not produce the same sequence of models.
  Example: $X_1, X_2 \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

  $$X_3 = X_1 + 3X_2 Y = X_1 + 2X_2 + \epsilon \tag{11}$$

  Regressing Y onto X_1, X_2, X_3:
  - Forward: $\{X_3\} \to \{X_3, X_2\} \to \{X_3, X_2, X_1\}$
  - Backward: $\{X_1, X_2, X_3\} \to \{X_1, X_2\} \to \{X_2\}$

Other stepwise selection methods

- *Mixed stepwise selection*: Do forward selection, but at every step, remove any variables that are no longer "necessary"
  - e.g. using p-values
- *Forward stagewise selection*: Do forward selection, but after every step, modify the remaining predictors such that they are uncorrelated to the selected predictors.
- etc.

Issues with stepwise methods

Important things to keep in mind:

- The selected model is not guaranteed to be optimal
  - There are often several equally good models
- The procedure does not take into account a researcher's knowledge about the predictors
- Outliers can have a large impact on the procedure
- Some predictors should be considered together as a group (e.g. dummy indicators for seasons of the year)
- The coefficients, R^2, p-values, CI's, etc are all biased/invalid
- Should not over-interpret the order that the predictors are included
- Cannot conclude that all variables included are important, or all excluded variables are unimportant

Shrinkage methods

Allows us to use all p predictors, but will regularize (i.e. shrink) their coefficients in some way.

- Common to shrink them towards 0

Shrinkage methods

Allows us to use all p predictors, but will regularize (i.e. shrink) their coefficients in some way.

- Common to shrink them towards 0

Question: Why would shrunk coefficients be better?

- Will introduce bias, but can significantly reduce the variance
    - If the variance is noticeably larger, this decreases the test error
- There are Bayesian motivations to do this: the prior tends to shrink the parameters.

Shrinkage methods

Allows us to use all p predictors, but will regularize (i.e. shrink) their coefficients in some way.

- Common to shrink them towards 0

Question: Why would shrunk coefficients be better?

- Will introduce bias, but can significantly reduce the variance
  - If the variance is noticeably larger, this decreases the test error
- There are Bayesian motivations to do this: the prior tends to shrink the parameters.

Three common shrinkage methods:

1. Ridge regression
2. Lasso regression
3. Elastic net

Ridge regression

Ridge regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{12}$$

In blue: the model RSS
In red: the squared ℓ_2 norm of β, or $\|\beta\|_2^2$

The parameter $\lambda > 0$ is a tuning parameter. It modulates the importance of fit vs. shrinkage.

- Typically determined via e.g. cross-validation

Ridge regression

Writing our loss function in matrix form

$$(\mathbf{Y} - \mathbf{X}\beta)^\top(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^\top\beta \tag{13}$$

it can be shown that

$$\hat{\beta}_n^{ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I_n})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{Y} \tag{14}$$

- So ridge regression simply adds a positive constant to $\mathbf{X}^\top\mathbf{X}$, making it non-singular.

Ridge regression

Under the linear model, the mean and covariance of $\hat{\beta}_n^{ridge}$ are:

$$\begin{aligned}\mathbb{E}[\hat{\beta}_n^{ridge}|\mathbf{X}] &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I_n})^{-\mathbf{1}}\mathbf{X}^\top\mathbb{E}[\mathbf{Y}|\mathbf{X}] \\ &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I_n})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X}\beta\end{aligned} \tag{15}$$

$$\begin{aligned}Cov[\hat{\beta}_n^{ridge}|\mathbf{X}] = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I_n})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{Cov}[\mathbf{Y}|\mathbf{X}] \\ X^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I_n})^{-\mathbf{1}} \\ = \sigma^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I_n})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X}^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I_n})^{-\mathbf{1}}\end{aligned} \tag{16}$$

Scaling predictors

In least-squares regression, scaling the variables has no effect on the fit of the model:

$$Y = X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{17}$$

e.g. Multiplying X_1 by c can be compensated by dividing $\hat{\beta}_1$ by c

- i.e. Doing this results in the same RSS

Scaling predictors

In least-squares regression, scaling the variables has no effect on the fit of the model:

$$Y = X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{17}$$

e.g. Multiplying X_1 by c can be compensated by dividing $\hat{\beta}_1$ by c

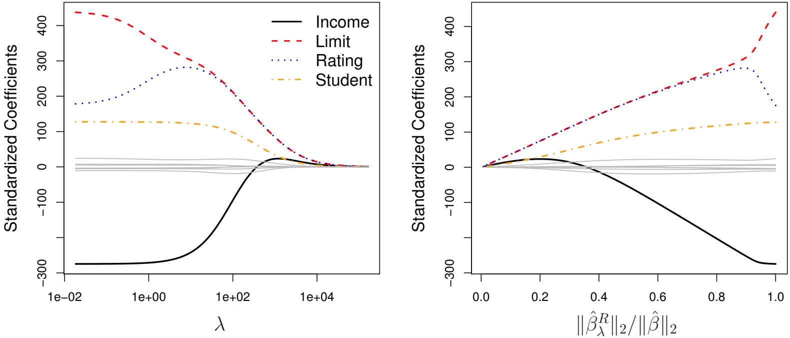
- i.e. Doing this results in the same RSS

This is not true for ridge regression!

- Due to $\|\beta\|_2^2$
- **In practice**: standardize all predictors (i.e. center and scale such that it has sample variance 1)
  - e.g. *glmnet* (by Hastie, Tibshirani, and Friedman)

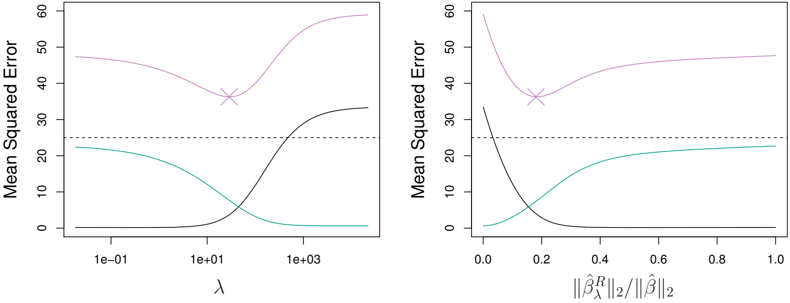
Example: Ridge regression

Ridge regression of `default` in the Credit dataset.



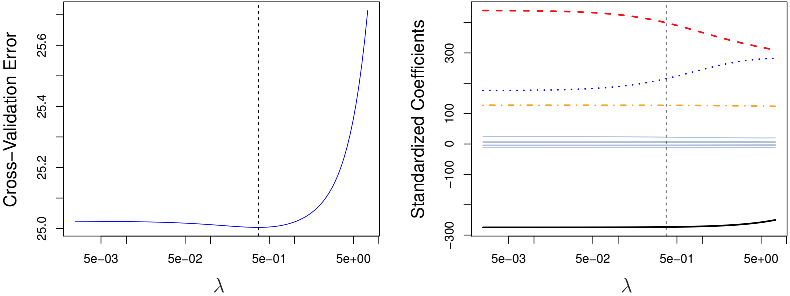
Bias variance trade-off

Computing the bias, variance, and test error as a function of λ (in simulation).



Cross validation would yield an estimate of the test error.

Selecting λ by cross-validation



The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{18}$$

In blue: the model RSS

In red: the ℓ_1 norm of β, or $\|\beta\|_1$

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{18}$$

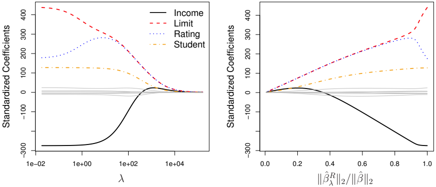
In blue: the model RSS

In red: the ℓ_1 norm of β, or $\|\beta\|_1$ **Note**: Unlike ridge regression, LASSO does not have a closed form solution.

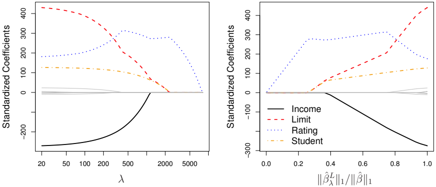
Why would we use the Lasso instead of Ridge regression?

- Ridge regression shrinks all the coefficients to a non-zero value
- The Lasso shrinks some of the coefficients all the way to zero.
  - Similar to subset selection: will select variables for you

Ridge regression of `default` in the Credit dataset.



Lasso regression of `default` in the Credit dataset.



An alternative formulation for regularization

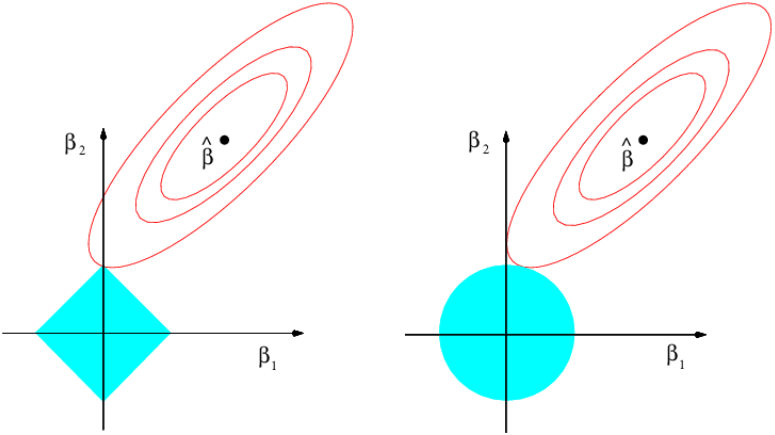
- **Ridge:** for every λ, there is an s such that $\hat{\beta}^R_\lambda$ solves:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j^2 < s \quad (19)$$

- **Lasso:** for every λ, there is an s such that $\hat{\beta}^L_\lambda$ solves:

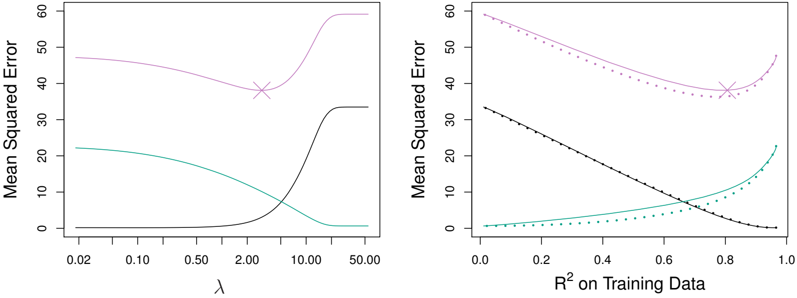
$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| < s \quad (20)$$

An alternative formulation for regularization



When is the Lasso better than Ridge?

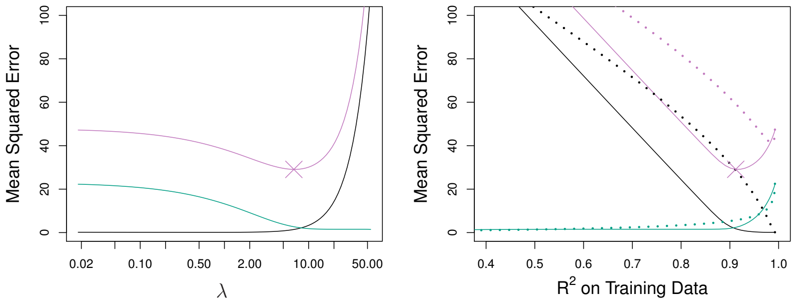
Example 1. Most of the coefficients are non-zero.



- Bias, Variance, MSE. The Lasso (—), Ridge $(\cdots)$.
- The bias is about the same for both methods.
- The variance of Ridge regression is smaller, so is the MSE.

When is the Lasso better than Ridge?

Example 2. Only 2 coefficients are non-zero.



- Bias, Variance, MSE. The Lasso (—), Ridge (· · ·).
- The bias, variance, and MSE are lower for the Lasso.

Elastic Net

Combines $\|\beta\|_2^2$ (ridge) and $\|\beta\|_1$ (lasso) penalties.

Elastic net solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \quad (21)$$

In blue: the model RSS
In red: both $\|\beta\|_2^2$ and $\|\beta\|_1$

This provides a nice trade off between sparsity and grouping.

Elastic Net

Combines $\|\beta\|_2^2$ (ridge) and $\|\beta\|_1$ (lasso) penalties.

Elastic net solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \quad (21)$$

In blue: the model RSS
In red: both $\|\beta\|_2^2$ and $\|\beta\|_1$

This provides a nice trade off between sparsity and grouping.

Typically, we define $\alpha = \frac{\lambda_2}{\lambda_2 + \lambda_1}$ and instead optimize:

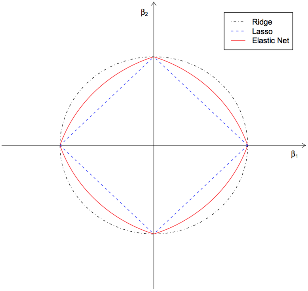
$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j| (22)$$

An alternative formulation for regularization

Elastic net:

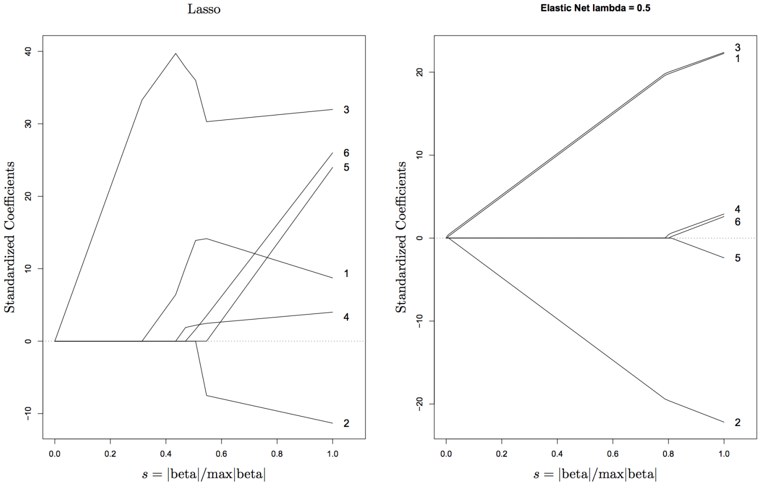
$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{i,j}\right)^2\right\} \text{ s.t. } \alpha\|\beta\|_2^2+(1-\alpha)\|\beta\|_1 < s (23)$$

2-dimensional illustration $\alpha = 0.5$



- Singularities at the vertexes (to encourage sparsity)
- Strict convex edges (to encourage grouping)
  - The strength of convexity varies with α

Example: Elastic net



Shrinkage summaries

$$\min_{\beta} \quad \sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{i,j}\right)^2+\lambda_2\sum_{j=1}^{p}\beta_j^2+\lambda_1\sum_{j=1}^{p}|\beta_j| \quad (24)$$

Method	Shrinkage parameters
OLS	$\lambda_1=\lambda_2=0$
Ridge	$\lambda_1=0, \lambda_2>0$
LASSO	$\lambda_1>0, \lambda_2=0$
Elastic net	$\lambda_1>0, \lambda_2>0$
$\hat{\beta}_n=0$	$\lambda_1=\infty$ or $\lambda_2=\infty$

A very special case

Suppose $n = p$, such that our predictor matrix is $X_{n \times n}$.
Then, the objective function in ridge regression can be simplified to

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{25}$$

and we can minimize the terms involving each β_j individually:

$$(y_j - \beta_j)^2 + \lambda \beta_j^2 \tag{26}$$

In this scenario, it can be shown that

$$\hat{\beta}_j^{ridge} = \frac{y_j}{1 + \lambda} \tag{27}$$

A very special case

Suppose $n = p$, such that our predictor matrix is $X_{n \times n}$.
We have a similar story for LASSO:

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{28}$$

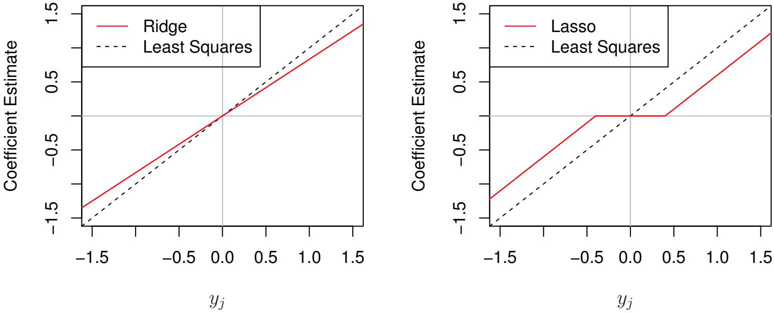
and we can minimize the terms involving each β_j individually:

$$(y_j - \beta_j)^2 + \lambda |\beta_j| \tag{29}$$

In this scenario, it can be shown that

$$\hat{\beta}_j^{LASSO} = \begin{cases} y_j - \lambda/2, & \text{if } y_j > \lambda/2 \\ y_j + \lambda/2, & \text{if } y_j < -\lambda/2 \\ 0, & \text{if } |y_j| < \lambda/2 \end{cases} \tag{30}$$

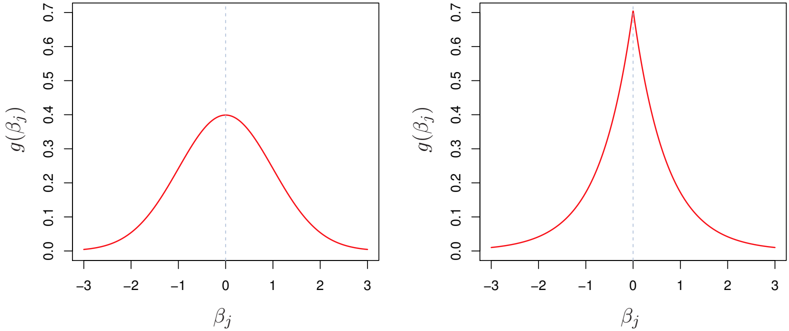
Lasso and Ridge coefficients as a function of λ



Bayesian interpretations

Ridge: $\hat{\beta}^R$ is the posterior mean, with a Normal prior on β.

Lasso: $\hat{\beta}^L$ is the posterior mode, with a Laplace prior on β.



Things to consider

- If desired, we could instead consider L_q penalties for values other than 0, 1, and 2 (e.g. $q \in (1, 2)$ or $q > 2$).
- Regularization methods such as the elastic net have been extended to generalized linear models (GLM) as well.
- L_1 and L_2 penalties are also used in contexts other than linear models (e.g. neural networks).
- As usual, we are faced with the bias-variance tradeoff when choosing our shrinkage parameters, λ_1 and λ_2.
- Other regularized methods are also available, e.g.
  - Non-negative Garotte Regression
  - Least Angle Regression
  - Best subset

Degrees of freedom

Degrees of freedom give us a measure of our model's complexity, i.e. the number of free parameters to fit on our data.

- For OLS, the degrees of freedom is equal to $p + 1$.
- In regularized regression, our parameters are estimated in a restricted manner, controlled by λ_1 and λ_2.
  - Effectively reduced the degrees of freedom in our model
- We can still compare across models using an *effective degrees of freedom*:

$$df(y, \hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} Cov[y_i, \hat{y}_i | x_i] \tag{31}$$

- In the case of OLS, this can be shown to reduce to the ''standard" degrees of freedom, i.e. $p + 1$.

References

[1] ISL. Chapters 6.

[2] ESL. Chapter 18.