CS189/CS289A
Introduction to Machine Learning
Lecture 10: Regression and Regularization

Peter Bartlett

February 19, 2015

Outline

Outline

- Review: Bias and variance

Outline

- Review: Bias and variance
- Subset selection

Outline

- Review: Bias and variance
- Subset selection
- Shrinkage:

Outline

- Review: Bias and variance
- Subset selection
- Shrinkage:
    - Ridge regression

Outline

- Review: Bias and variance
- Subset selection
- Shrinkage:
    - Ridge regression
    - Lasso

Outline

- **Review: Bias and variance**
- Subset selection
- Shrinkage:
  - Ridge regression
  - Lasso

Review: Bias and variance

Review: Bias and variance

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$.

Review: Bias and variance

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.

Review: Bias and variance

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.
- We can decompose the excess risk:

Review: Bias and variance

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.
- We can decompose the excess risk:

$$\mathbb{E}R(\hat{f}) - R^* = \mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right]$$

Review: Bias and variance

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.
- We can decompose the excess risk:

$$\begin{aligned}
\mathbb{E}R(\hat{f}) - R^* &= \mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\hat{f}(X) + \mathbb{E}\hat{f}(X) - f^*(X)\right)^2\right]
\end{aligned}$$

Review: Bias and variance

- We use randomly chosen training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a prediction rule $\hat{f}$. So that prediction rule is random, and its risk $R(\hat{f})$ is a random variable.
- We can decompose the excess risk:

$$
\begin{aligned}
\mathbb{E}R(\hat{f}) - R^* &= \mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\hat{f}(X) + \mathbb{E}\hat{f}(X) - f^*(X)\right)^2\right] \\
&= \underbrace{\mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\hat{f}(X)\right)^2\right]}_{\text{variance}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}\hat{f}(X) - f^*(X)\right)^2\right]}_{\text{bias}^2}.
\end{aligned}
$$

Trading off bias and variance

Some increase in bias (i.e., decreased complexity) can give a big decrease in variance.

Trading off bias and variance

Some increase in bias (i.e., decreased complexity) can give a big decrease in variance.
For instance, suppose we have a vector of covariates corresponding to polynomials of degree 15 in a scalar variable x.

Trading off bias and variance

Some increase in bias (i.e., decreased complexity) can give a big decrease in variance.
For instance, suppose we have a vector of covariates corresponding to polynomials of degree 15 in a scalar variable x.

- Including more of these covariates (i.e., richer polynomials) will give lower bias.

Trading off bias and variance

Some increase in bias (i.e., decreased complexity) can give a big decrease in variance.
For instance, suppose we have a vector of covariates corresponding to polynomials of degree 15 in a scalar variable x.

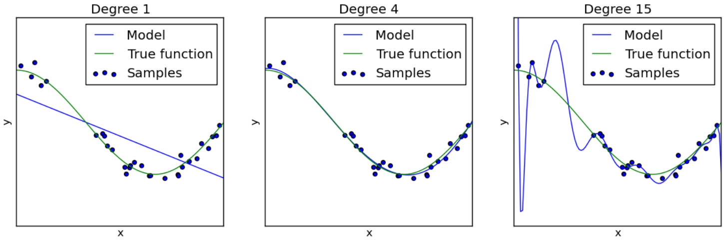
- Including more of these covariates (i.e., richer polynomials) will give lower bias.
- However, for a fixed sample size, the variance will be larger.

Trading off bias and variance

Some increase in bias (i.e., decreased complexity) can give a big decrease in variance.

For instance, suppose we have a vector of covariates corresponding to polynomials of degree 15 in a scalar variable x.

- Including more of these covariates (i.e., richer polynomials) will give lower bias.
- However, for a fixed sample size, the variance will be larger.



Trading off bias and variance

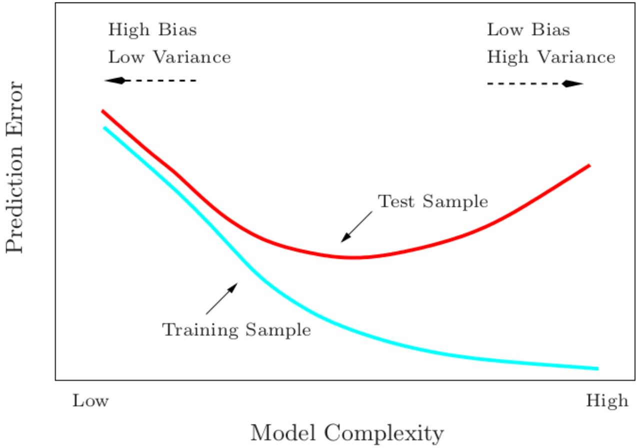


FIGURE 2.11. *Test and training error as a function of model complexity.*

Trading off bias and variance

Model selection

How do we choose an appropriate model complexity?

Trading off bias and variance

Model selection

How do we choose an appropriate model complexity?

For a linear model, the complexity depends on:

Trading off bias and variance

Model selection

How do we choose an appropriate model complexity?

For a linear model, the complexity depends on:

- The number of predictor variables.

Trading off bias and variance

Model selection

How do we choose an appropriate model complexity?

For a linear model, the complexity depends on:

- The number of predictor variables.
  Equivalently, the number of non-zero coefficients of $\hat{\beta}$.

Trading off bias and variance

Model selection

How do we choose an appropriate model complexity?

For a linear model, the complexity depends on:

- The number of predictor variables.
  Equivalently, the number of non-zero coefficients of $\hat{\beta}$.
- More subtly, the size of coefficients of $\hat{\beta}$.

Trading off bias and variance

Model selection

How do we choose an appropriate model complexity?

For a linear model, the complexity depends on:

- The number of predictor variables.
  Equivalently, the number of non-zero coefficients of $\hat{\beta}$.
- More subtly, the size of coefficients of $\hat{\beta}$.

Restricting complexity corresponds to increasing bias and decreasing variance.

Outline

- Review: Bias and variance
- **Subset selection**
- Shrinkage:
    - Ridge regression
    - Lasso

Subset selection

- Consider the 2^p different subsets of the variables.

Subset selection

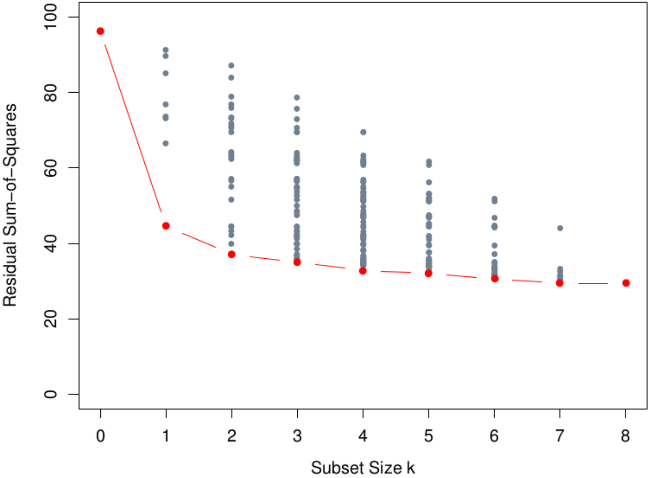
- Consider the 2^p different subsets of the variables.
- Fit a linear model with each subset.

Subset selection

- Consider the 2^p different subsets of the variables.
- Fit a linear model with each subset.
- Decide which of these linear models to use.

Subset selection

- Consider the 2^p different subsets of the variables.
- Fit a linear model with each subset.
- Decide which of these linear models to use.



Subset selection

Subset selection

- RSS decreases as the complexity increases (because the best fit with a smaller subset is always possible with a larger subset).

Subset selection

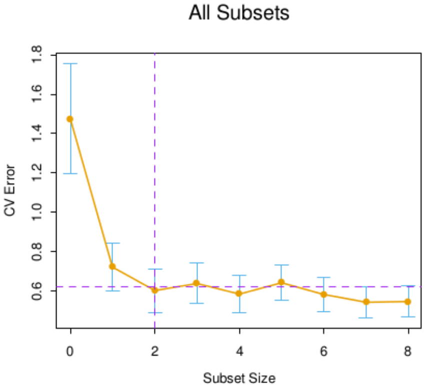
- RSS decreases as the complexity increases (because the best fit with a smaller subset is always possible with a larger subset).
- How do we decide which model complexity to use?

Subset selection

- RSS decreases as the complexity increases (because the best fit with a smaller subset is always possible with a larger subset).
- How do we decide which model complexity to use?
- **Cross-validation.**

Subset selection

- RSS decreases as the complexity increases (because the best fit with a smaller subset is always possible with a larger subset).
- How do we decide which model complexity to use?
- **Cross-validation.**



Subset selection: Computational shortcuts

Subset selection: Computational shortcuts

- There are lots of subsets of p variables.

Subset selection: Computational shortcuts

- There are lots of subsets of p variables.
- Rather than trying all of them, we might hope to find a path through subset space that visits good subsets.

Subset selection: Computational shortcuts

- There are lots of subsets of p variables.
- Rather than trying all of them, we might hope to find a path through subset space that visits good subsets.

Forward-stepwise selection

Subset selection: Computational shortcuts

- There are lots of subsets of p variables.
- Rather than trying all of them, we might hope to find a path through subset space that visits good subsets.

Forward-stepwise selection

$\{x_0\}$

Subset selection: Computational shortcuts

- There are lots of subsets of p variables.
- Rather than trying all of them, we might hope to find a path through subset space that visits good subsets.

Forward-stepwise selection

$$\{x_0\}$$
$$\downarrow$$
$$\{x_0, x_3\}$$

Subset selection: Computational shortcuts

- There are lots of subsets of p variables.
- Rather than trying all of them, we might hope to find a path through subset space that visits good subsets.

Forward-stepwise selection

$$\{x_0\}$$
$$\downarrow$$
$$\{x_0, x_3\}$$
$$\downarrow$$
$$\{x_0, x_3, x_{17}\}$$

Subset selection: Computational shortcuts

- There are lots of subsets of p variables.
- Rather than trying all of them, we might hope to find a path through subset space that visits good subsets.

Forward-stepwise selection

$$\{x_0\}$$
$$\downarrow$$
$$\{x_0, x_3\}$$
$$\downarrow$$
$$\{x_0, x_3, x_{17}\}$$
$$\downarrow$$
$$\{x_0, x_3, x_{17}, x_5\}$$

Subset selection: Computational shortcuts

- There are lots of subsets of p variables.
- Rather than trying all of them, we might hope to find a path through subset space that visits good subsets.

Forward-stepwise selection

$$\{x_0\}$$
$$\downarrow$$
$$\{x_0, x_3\}$$
$$\downarrow$$
$$\{x_0, x_3, x_{17}\}$$
$$\downarrow$$
$$\{x_0, x_3, x_{17}, x_5\}$$
$$\downarrow$$
$$\{x_0, x_3, x_{17}, x_5, x_1\}$$

Subset selection: Computational shortcuts

- There are lots of subsets of p variables.
- Rather than trying all of them, we might hope to find a path through subset space that visits good subsets.

Forward-stepwise selection

$$
\begin{array}{l}
\{x_0\} \\
\downarrow \\
\{x_0, x_3\} \\
\downarrow \\
\{x_0, x_3, x_{17}\} \\
\downarrow \\
\{x_0, x_3, x_{17}, x_5\} \\
\downarrow \\
\{x_0, x_3, x_{17}, x_5, x_1\} \\
\downarrow \\
\{x_0, x_3, x_{17}, x_5, x_1, x_{12}\} \\
\downarrow \\
\vdots
\end{array}
$$

Subset selection: Computational shortcuts

Backward-stepwise selection

Subset selection: Computational shortcuts

Backward-stepwise selection

$\{x_0, x_1, x_2, x_3, x_4, x_5, x_6\}$ (Need $n > p$)

Subset selection: Computational shortcuts

Backward-stepwise selection

$$\{x_0, x_1, x_2, x_3, x_4, x_5, x_6\} \qquad \text{(Need } n > p\text{)}$$
$$\downarrow$$
$$\{x_0, x_1, x_2, x_4, x_5, x_6\}$$

Subset selection: Computational shortcuts

Backward-stepwise selection

$$\{x_0, x_1, x_2, x_3, x_4, x_5, x_6\} \qquad \text{(Need } n > p\text{)}$$
$$\downarrow$$
$$\{x_0, x_1, x_2, x_4, x_5, x_6\}$$
$$\downarrow$$
$$\{x_0, x_1, x_2, x_4, x_5\}$$

Subset selection: Computational shortcuts

Backward-stepwise selection

$$\{x_0, x_1, x_2, x_3, x_4, x_5, x_6\} \qquad \text{(Need } n > p\text{)}$$

$$\downarrow$$

$$\{x_0, x_1, x_2, x_4, x_5, x_6\}$$

$$\downarrow$$

$$\{x_0, x_1, x_2, x_4, x_5\}$$

$$\downarrow$$

$$\{x_0, x_2, x_4, x_5\}$$

Subset selection: Computational shortcuts

Backward-stepwise selection

$$
\begin{array}{ll}
\{x_0, x_1, x_2, x_3, x_4, x_5, x_6\} & \qquad \text{(Need } n > p\text{)} \\
\downarrow & \\
\{x_0, x_1, x_2, x_4, x_5, x_6\} & \\
\downarrow & \\
\{x_0, x_1, x_2, x_4, x_5\} & \\
\downarrow & \\
\{x_0, x_2, x_4, x_5\} & \\
\downarrow & \\
\{x_0, x_4, x_5\} &
\end{array}
$$

Subset selection: Computational shortcuts

Backward-stepwise selection

$$\begin{array}{ll}
\{x_0, x_1, x_2, x_3, x_4, x_5, x_6\} & \qquad \text{(Need } n > p\text{)} \\
\downarrow & \\
\{x_0, x_1, x_2, x_4, x_5, x_6\} & \\
\downarrow & \\
\{x_0, x_1, x_2, x_4, x_5\} & \\
\downarrow & \\
\{x_0, x_2, x_4, x_5\} & \\
\downarrow & \\
\{x_0, x_4, x_5\} & \\
\downarrow & \\
\{x_0, x_4\} &
\end{array}$$

Subset selection: Computational shortcuts

Backward-stepwise selection

$\{x_0, x_1, x_2, x_3, x_4, x_5, x_6\}$ (Need $n > p$)

$\downarrow$

$\{x_0, x_1, x_2, x_4, x_5, x_6\}$

$\downarrow$

$\{x_0, x_1, x_2, x_4, x_5\}$

$\downarrow$

$\{x_0, x_2, x_4, x_5\}$

$\downarrow$

$\{x_0, x_4, x_5\}$

$\downarrow$

$\{x_0, x_4\}$

$\downarrow$

$\{x_0\}$

Outline

- Review: Bias and variance
- Subset selection
- **Shrinkage:**
  - Ridge regression
  - Lasso

Shrinkage methods

Model selection

How do we choose an appropriate model complexity?
For a linear model, the complexity depends on:

- The number of predictor variables.
  Equivalently, the number of non-zero coefficients of $\hat{\beta}$.
- More subtly, **the size of coefficients of $\hat{\beta}$.**

Model selection

How do we choose an appropriate model complexity?
For a linear model, the complexity depends on:

- The number of predictor variables.
  Equivalently, the number of non-zero coefficients of $\hat{\beta}$.
- More subtly, **the size of coefficients of $\hat{\beta}$.**

- Subset selection: choose a subset of variables that will have non-zero coefficients.

Shrinkage methods

Model selection

How do we choose an appropriate model complexity?
For a linear model, the complexity depends on:

- The number of predictor variables.
  Equivalently, the number of non-zero coefficients of $\hat{\beta}$.
- More subtly, **the size of coefficients of** $\hat{\beta}$**.**

- Subset selection: choose a subset of variables that will have non-zero coefficients.
- Shrinkage: encourage the linear predictor's coefficients to be small.

Shrinkage methods: Ridge regression

Shrinkage methods: Ridge regression

To encourage the parameters β of our linear prediction rule $x \mapsto x'\beta$ to be small, we can modify the least squares criterion:

Shrinkage methods: Ridge regression

To encourage the parameters β of our linear prediction rule $x \mapsto x'\beta$ to be small, we can modify the least squares criterion:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 \qquad\qquad \right),$$

Shrinkage methods: Ridge regression

To encourage the parameters β of our linear prediction rule $x \mapsto x'\beta$ to be small, we can modify the least squares criterion: either punish large values through a penalty term:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 \qquad\qquad \right),$$

Shrinkage methods: Ridge regression

To encourage the parameters β of our linear prediction rule $x \mapsto x'\beta$ to be small, we can modify the least squares criterion: either punish large values through a penalty term:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right),$$

Shrinkage methods: Ridge regression

To encourage the parameters β of our linear prediction rule $x \mapsto x'\beta$ to be small, we can modify the least squares criterion: either punish large values through a penalty term:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right),$$

or forbid them through a constraint:

Shrinkage methods: Ridge regression

To encourage the parameters β of our linear prediction rule $x \mapsto x'\beta$ to be small, we can modify the least squares criterion: either punish large values through a penalty term:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right),$$

or forbid them through a constraint:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2$$

Shrinkage methods: Ridge regression

To encourage the parameters β of our linear prediction rule $x \mapsto x'\beta$ to be small, we can modify the least squares criterion: either punish large values through a penalty term:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right),$$

or forbid them through a constraint:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \leq B.$$

Shrinkage methods: Ridge regression

To encourage the parameters β of our linear prediction rule $x \mapsto x'\beta$ to be small, we can modify the least squares criterion: either punish large values through a penalty term:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right),$$

or forbid them through a constraint:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2$$

$$\text{s.t. } \sum_{j=1}^{p} \beta_j^2 \leq B.$$

(These problems are equivalent, in the sense that for given data, for every value of λ, there is a B such that the solutions are identical.)

Shrinkage methods: Ridge regression

- It is standard not to penalize the β_0 associated with the offset $x_0 = 1$.

Shrinkage methods: Ridge regression

- It is standard not to penalize the β_0 associated with the offset $x_0 = 1$.
- We can write the penalized RSS as

$$\|y - X\beta\|^2 + \lambda\|\beta\|^2,$$

Shrinkage methods: Ridge regression

- It is standard not to penalize the β_0 associated with the offset $x_0 = 1$.
- We can write the penalized RSS as

$$\|y - X\beta\|^2 + \lambda\|\beta\|^2,$$

  which is minimized by

$$\hat{\beta}^{\text{ridge}} = \left(X'X + \lambda I\right)^{-1} X'y.$$

Shrinkage methods: Ridge regression

- It is standard not to penalize the β_0 associated with the offset $x_0 = 1$.
- We can write the penalized RSS as

$$\|y - X\beta\|^2 + \lambda\|\beta\|^2,$$

  which is minimized by

$$\hat{\beta}^{\text{ridge}} = \left(X'X + \lambda I\right)^{-1} X'y.$$

- Notice that, even if $X'X$ is singular, this penalty ensures that there is a unique solution $\hat{\beta}$.

Shrinkage methods: Ridge regression

- It is standard not to penalize the β_0 associated with the offset $x_0 = 1$.
- We can write the penalized RSS as

$$\|y - X\beta\|^2 + \lambda\|\beta\|^2,$$

which is minimized by

$$\hat{\beta}^{\text{ridge}} = \left(X'X + \lambda I\right)^{-1} X'y.$$

- Notice that, even if $X'X$ is singular, this penalty ensures that there is a unique solution $\hat{\beta}$.
(If $X'X$ is singular, the covariates are linearly dependent, so changing β in some direction will not change the linear map on the space spanned by the data. Adding the penalty eliminates these cancellations.)

Shrinkage methods: Ridge regression

- It is standard not to penalize the β_0 associated with the offset $x_0 = 1$.
- We can write the penalized RSS as

$$\|y - X\beta\|^2 + \lambda\|\beta\|^2,$$

which is minimized by

$$\hat{\beta}^{\text{ridge}} = \left(X'X + \lambda I\right)^{-1} X'y.$$

- Notice that, even if $X'X$ is singular, this penalty ensures that there is a unique solution $\hat{\beta}$.
  (If $X'X$ is singular, the covariates are linearly dependent, so changing β in some direction will not change the linear map on the space spanned by the data. Adding the penalty eliminates these cancellations.)
- The solution depends on scaling of the covariates!
  It is common to standardize covariates (scale so variance is 1).

Shrinkage methods: Ridge regression

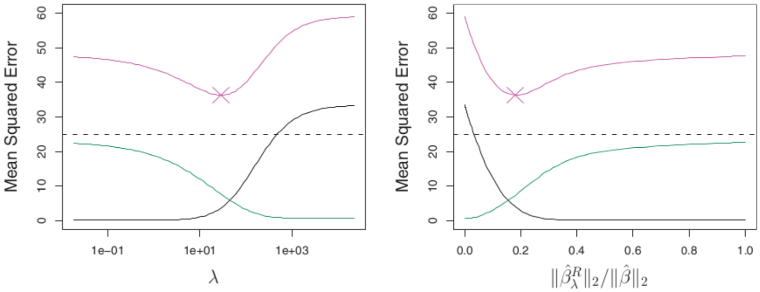


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Outline

- Review: Bias and variance
- Subset selection
- Shrinkage:
    - Ridge regression
    - **Lasso**

Shrinkage methods: Lasso

Just like ridge regression, except we replace the 2-norm of β with the 1-norm:

Shrinkage methods: Lasso

Just like ridge regression, except we replace the 2-norm of β with the 1-norm:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 \qquad\qquad \right),$$

Shrinkage methods: Lasso

Just like ridge regression, except we replace the 2-norm of β with the 1-norm:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

Shrinkage methods: Lasso

Just like ridge regression, except we replace the 2-norm of β with the 1-norm:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

or

Shrinkage methods: Lasso

Just like ridge regression, except we replace the 2-norm of β with the 1-norm:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

or

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

Shrinkage methods: Lasso

Just like ridge regression, except we replace the 2-norm of β with the 1-norm:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

or

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq B.$$

Shrinkage methods: Lasso

Just like ridge regression, except we replace the 2-norm of β with the 1-norm:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

or

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

$$\text{s.t. } \sum_{j=1}^{p} |\beta_j| \leq B.$$

(As before, these problems are equivalent.)

Shrinkage methods: Lasso

While ridge regression leads to reduced, but non-zero values of the coefficients, the Lasso forces some coefficients to be zero:

Shrinkage methods: Lasso

While ridge regression leads to reduced, but non-zero values of the coefficients, the Lasso forces some coefficients to be zero:

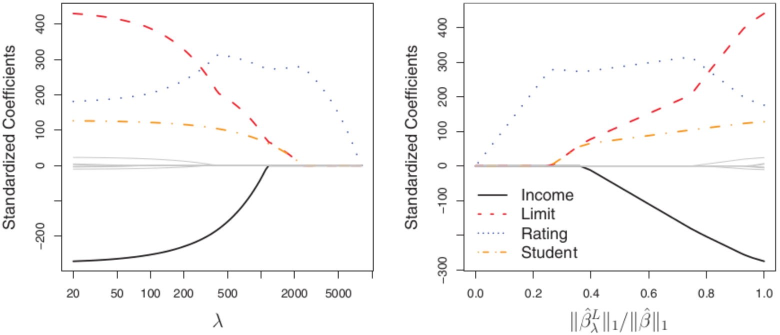


FIGURE 6.6. *The standardized lasso coefficients on the* Credit *data set are shown as a function of λ and $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.*

Shrinkage methods: Lasso

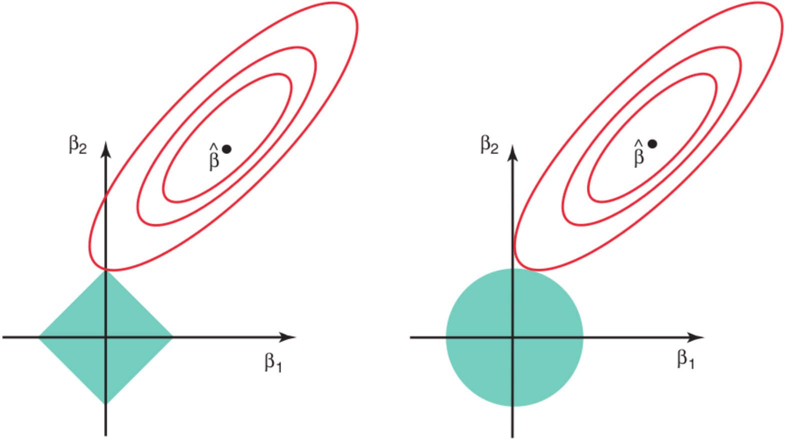
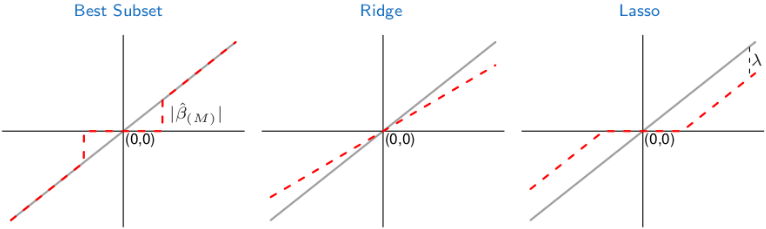


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Shrinkage methods: Lasso

TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of $\mathbf{X}$. M and λ are constants chosen by the corresponding techniques;* sign *denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



Outline

- Review: Bias and variance
- Subset selection
- Shrinkage:
    - Ridge regression
    - Lasso