Linear Model Selection and Regularization

Dr. Lan Du

Faculty of Information Technology, Monash University, Australia

FIT5149 week 6

Outline

Improve linear model fitting procedure

- Recall the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- The linear model has distinct advantages in terms of its interpretability and often shows good predictive performance, while the assumptions are satisfied.
- Improve the simple linear model:
  - replace ordinary least squares fitting with some alternative fitting procedures.
- Yield better prediction accuracy and model interpretability

Why consider alternatives to least squares?

- Prediction Accuracy: especially when p > n, to control the variance.
  - If the true linear relationship between the Y and X $\Rightarrow$ low bias
  - If $n \gg p \Rightarrow$ low variance
  - If $n \approx p \Rightarrow$ high variance & overfitting & poor prediction
  - if $n < p \Rightarrow$ infinite variance & no unique OLS coefficient estimate
- Model Interpretability:
  - When we have a large number of variables X, there will generally be some or many that are not associated with the response Y.
  - Including irrelevant variables leads to unnecessary complexity in the model
  - Removing irrelevant variables by setting their coefficient to 0 increases the interpretability of the resulting model.
- Solution: feature (variable) selection.

Three classes of selection methods

- Subsect Selection: Identifying a subset of all p predictors X that we believe to be related to the response Y, and then fitting the model least squares on the reduced set of variables.
  - Best subset selection
  - Forward/Backward stepwise selection
  - Hybrid selection
- Shrinkage, also known as regularisation
  - The estimated coefficients are shrunken towards zero relative to the least squares estimates.
  - The shrinkage has the effect of reducing variance.
  - The shrinkage can also perform variable selection.
    - Ridge regression: L2 regularisation
    - The Lasso: L1 regularisation
    - Elastic Net: the mixture of L1 and L2
    - Group Lasso
- Dimension Reduction: Involves projecting all p predictors into an M-dimensional space where $M < p$, and then fitting regression model.
  - e.g., Principle Component Analysis

Outline

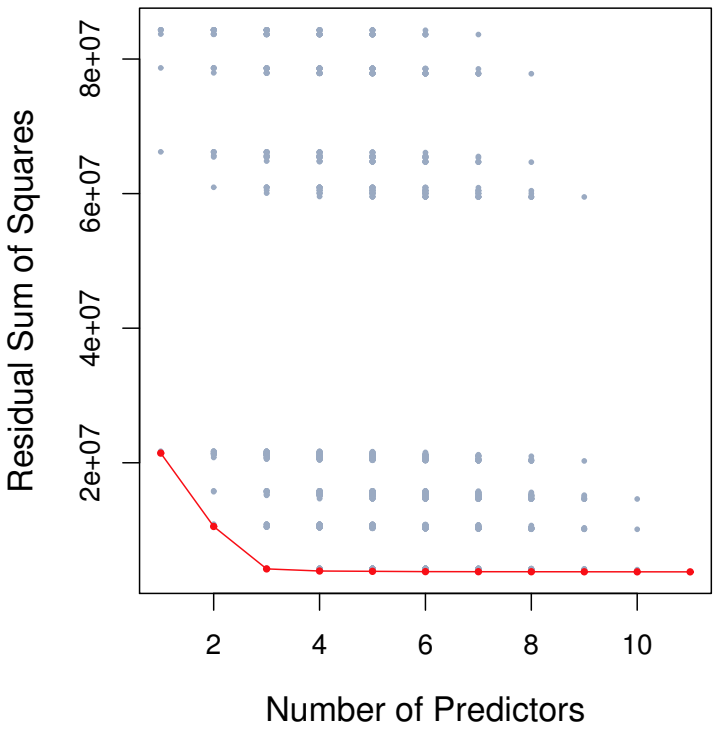
Best Subset Selection

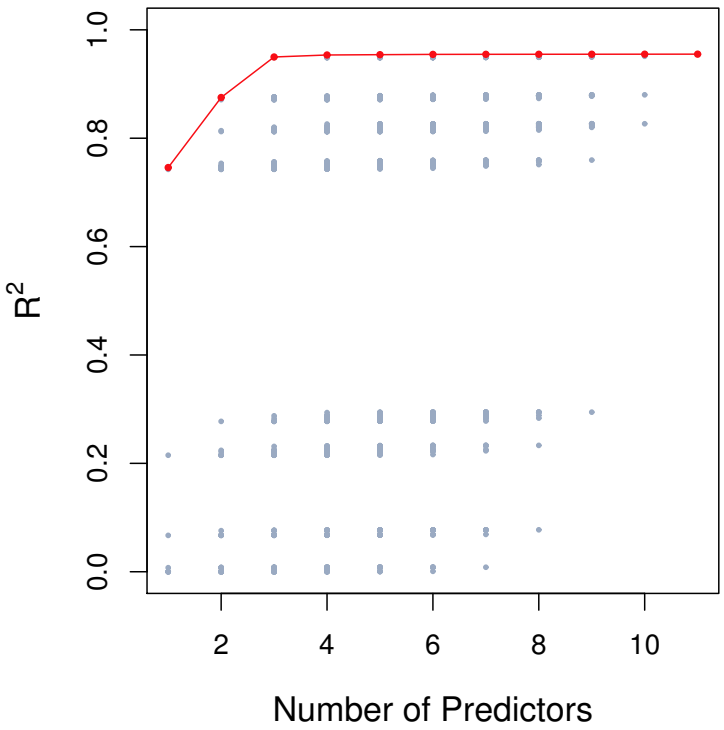
MONASH University

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest 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.

Example: Credit data set





For each possible model containing a subset of the ten predictors in the Credit data set, the RSS and R^2 are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and R^2. Though the data set contains only ten predictors, the x-axis ranges from 1 to 11, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables

More on Best Subset Selection

- It apply to other types of models, such as logistic regression.
  - The deviance — negative two times the maximized log-likelihood — plays the role of RSS for a broader class of models.
- For computational reasons, best subset selection cannot be applied with very large p. Why not?
- Overfitting: when p is large, larger the search space, the higher the chance of finding models with a low training error, which by no means guarantees a low test error.
- Stepwise methods: which explore a far more restricted set of models, are attractive alternatives to best subset selection.

Forward Stepwise Selection

MONASH University

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.

More on Forward Stepwise Selection

MONASH University

- Computational advantage over best subset selection is clear.
  - Best subset selection: 2^p models
  - Forward stepwise selection: $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.
  - Why not? Give an example.

More on Forward Stepwise Selection

MONASH University

- Computational advantage over best subset selection is clear.
  - Best subset selection: 2^p models
  - Forward stepwise selection: $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.
  - Why not? Give an example.
  - suppose that in a given data set with $p = 3$ predictors,
    - the best possible one-variable model contains X_1
    - the best possible two-variable model instead contains X_2 and X_3

    Then forward stepwise selection will fail to select the best possible two-variable model, because M_1 will contain X_1, so M_2 must also contain X_1 together with one additional variable.

More on Forward Stepwise Selection

MONASH University

- Computational advantage over best subset selection is clear.
  - Best subset selection: 2^p models
  - Forward stepwise selection: $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.
  - Why not? Give an example.
  - suppose that in a given data set with $p = 3$ predictors,
    - the best possible one-variable model contains X_1
    - the best possible two-variable model instead contains X_2 and X_3

    Then forward stepwise selection will fail to select the best possible two-variable model, because M_1 will contain X_1, so M_2 must also contain X_1 together with one additional variable.
- Forward stepwise selection can be applied even in the high-dimensional setting where n < p.
  - Just construct submodels $M_0, \ldots, M_{n-1}$ only. Why?

Compare best subset selection with forward selection

MONASH University

# 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

MONASH University

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.

More on Backward Stepwise Selection

MONASH University

- Like forward stepwise selection, the backward selection approach searches through only $1 + p(p+1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection.
- Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors.
- Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit).

Choosing the Optimal Model

MONASH University

- RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.
  - The RSS of these $p + 1$ models decreases monotonically, and the R^2 increases monotonically, as the number of features included in the models increases.
  - These quantities are related to the training error. Recall that training error is usually a poor estimate of test error.

  We wish to choose a model with low test error, not a model with low training error.
- How to estimate test error?

Choosing the Optimal Model

MONASH University

- RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.
  - The RSS of these $p + 1$ models decreases monotonically, and the R^2 increases monotonically, as the number of features included in the models increases.
  - These quantities are related to the training error. Recall that training error is usually a poor estimate of test error.

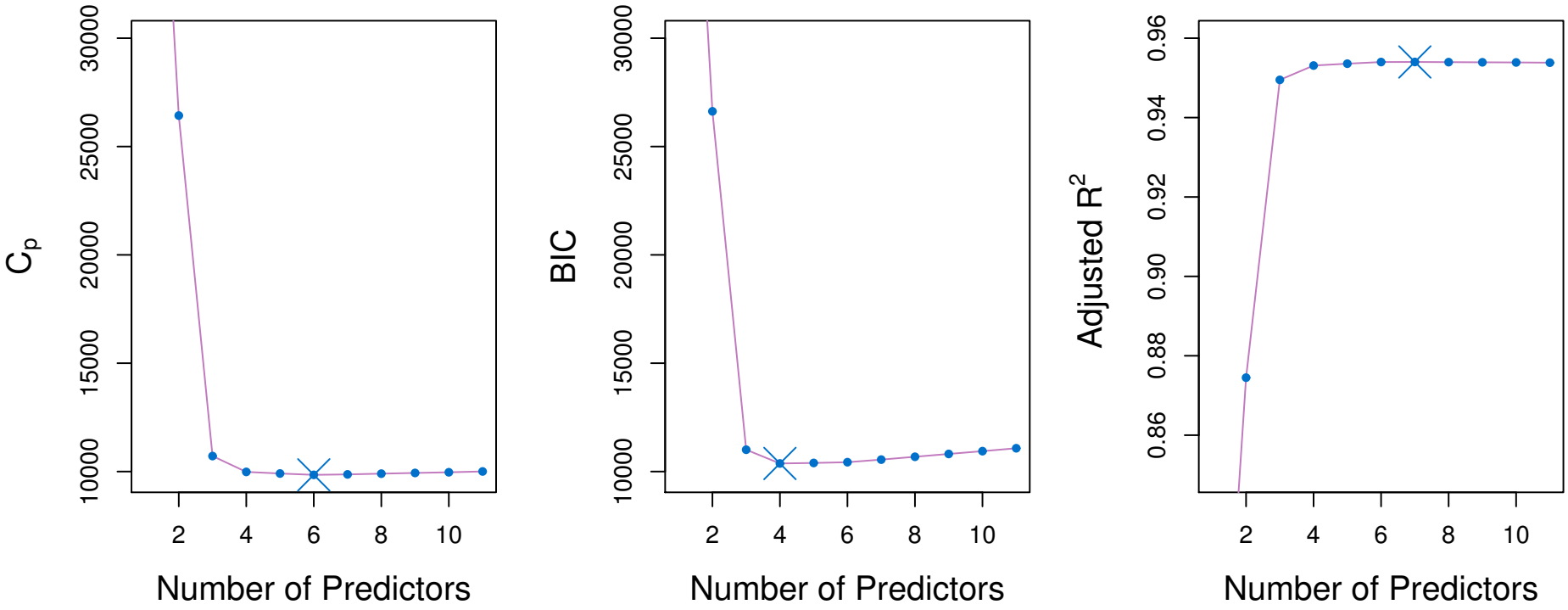
  We wish to choose a model with low test error, not a model with low training error.
- How to estimate test error?
  - To estimate test error by making an adjustment to the training error to account for the bias due to overfitting.
  - To directly estimate the test error, using either a validation set approach or a cross-validation approach.

Measures for selection the best model

MONASH University

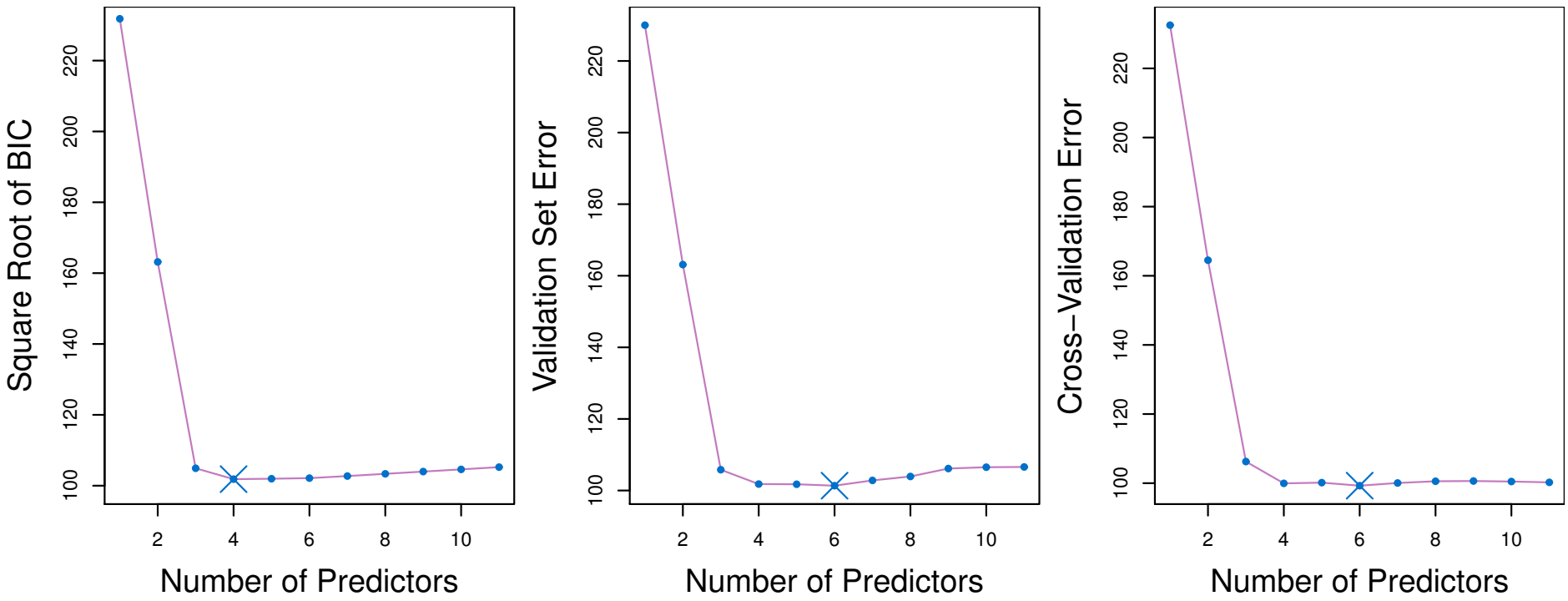
- Other measures can be used to select among a set of models with different numbers of variables:
  - Mallow's C_p: $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$
  - AIC (Akaike information criterion): $AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$
  - BIC (Bayesian information criterion): $BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)$
  - Adjusted R^2: $Adjusted\_R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$
- These methods add penalty to RSS for the number of variables (i.e. complexity) in the model.

Example on the Credit data



- A small value of C_p and BIC indicates a low error, and thus a better model.
- A large value for the Adjusted R^2 indicates a better model.

Validation and Cross-Validation



- We can compute the validation set error or the cross-validation error for each model under consideration, and then select the model for which the resulting estimated test error is smallest
- Advantage: it provides a direct estimate of the test error, and makes fewer assumptions about the true underlying model.

Outline

Shrinkage Methods

MONASH University

Ridge regression and Lasso

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

Ridge regression

- The Ordinary Least Squares (OLS) fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2$$

- The ridge regression coefficient estimates are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 + \lambda \parallel \boldsymbol{\beta} \parallel_2^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

where

  - $\lambda \geq 0$ is a regularisation parameter (or tuning parameter).
  - $\lambda \parallel \boldsymbol{\beta} \parallel_2^2$ is called a shrinkage penalty.

What does the shrinkage penalty do?

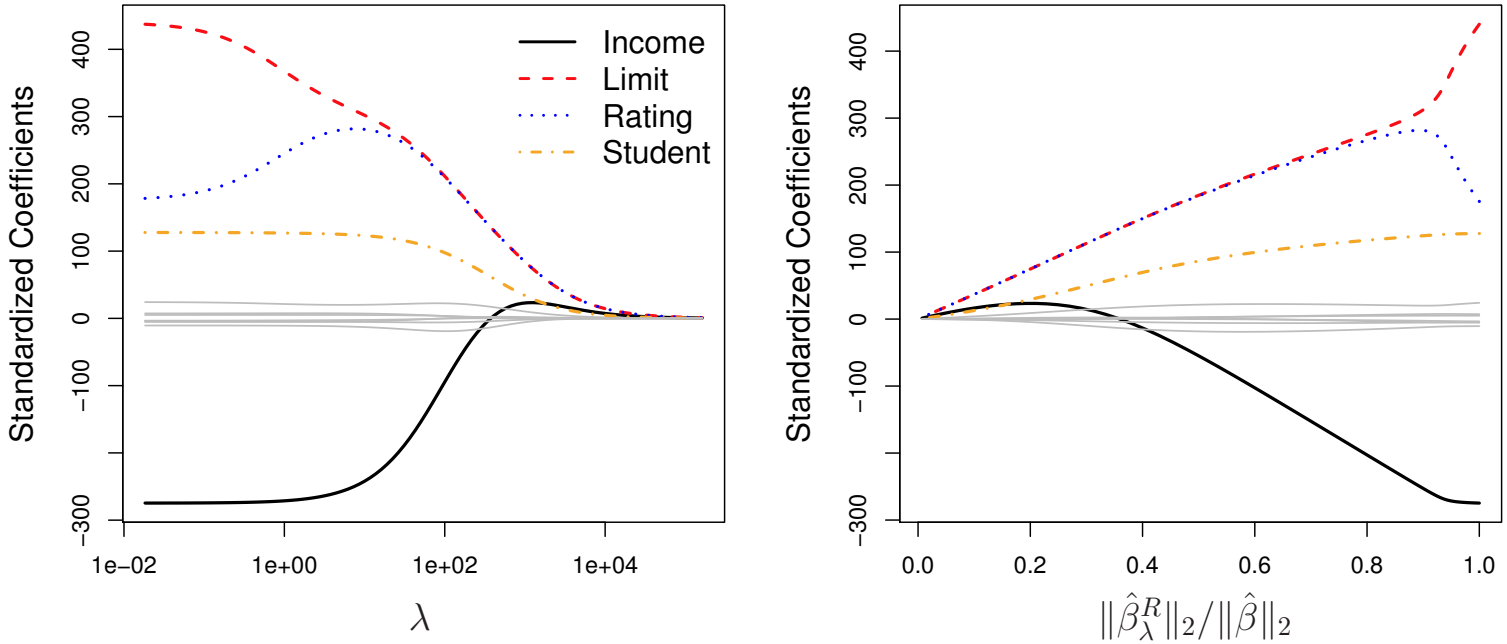


Figure: The standardized ridge regression coefficients are displayed for the Credit data set, as a function of λ and $\| \hat{\beta}^R_\lambda \|_2 / \| \hat{\beta} \|_2$

- The shrinkage penalty has the effect of shrinking the estimates of β_j towards zero.
  - $\lambda = 0$: ridge regression will produce the least squares estimates.
  - $\lambda \to \infty$: the ridge regression coefficient estimates will approach zero.

What does the shrinkage penalty do?

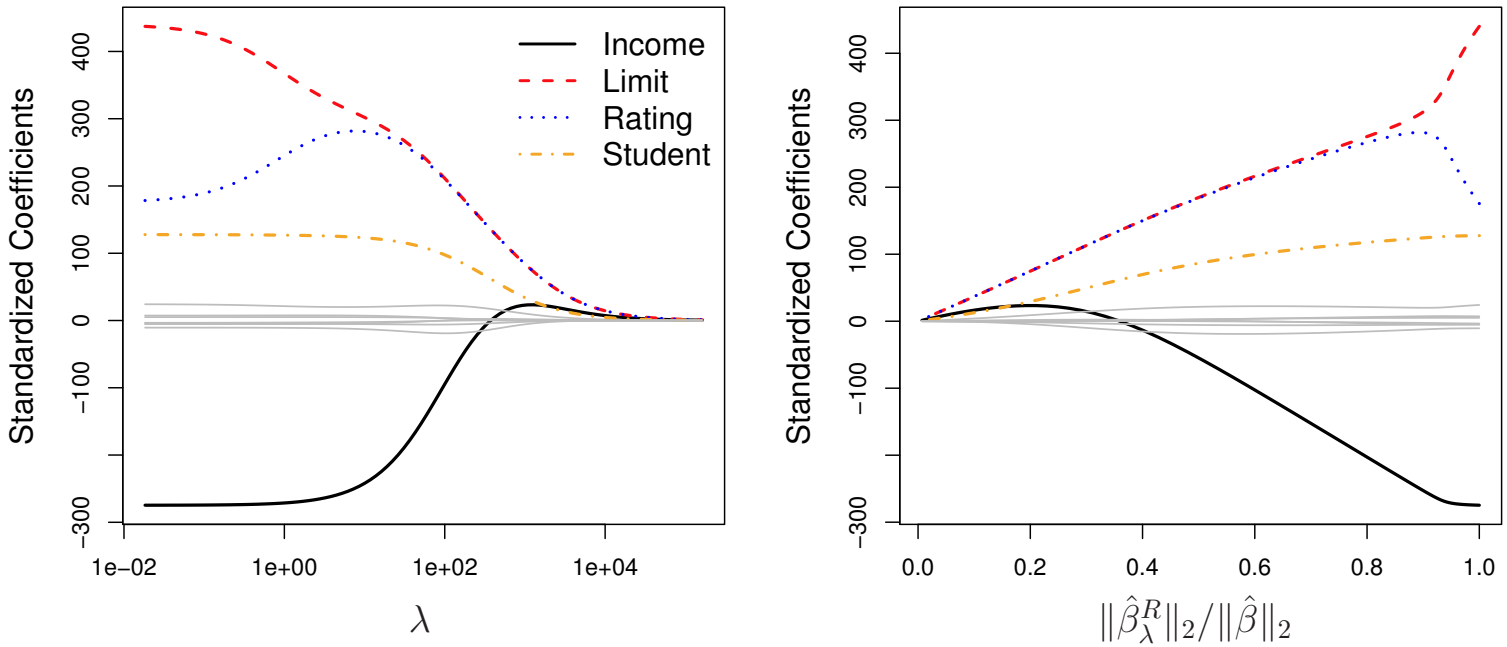


Figure: The standardized ridge regression coefficients are displayed for the Credit data set, as a function of λ and $\| \hat{\beta}^R_\lambda \|_2 / \| \hat{\beta} \|_2$

- The notation $\| \beta \|_2$ denotes the l_2 norm, $\| \beta \|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$, which measures the distance of β from zero.

Why does ridge regression improve over OLS?

Recall that MSE is a function of the variance plus the squared bias.

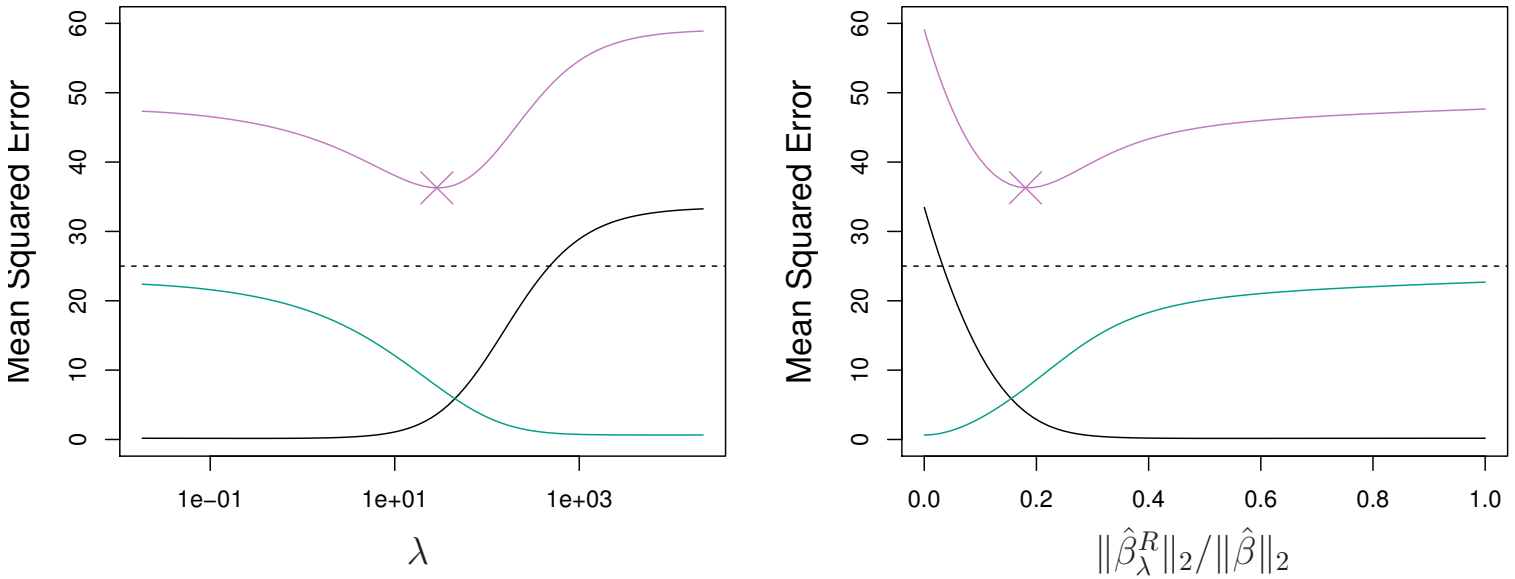


Figure: The Bias-Variance tradeoff with a simulated data set containing $p = 45$ predictors and $n = 50$ samples. It shows that as λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

- Black: Squared bias
- Green: Variance
- Purple: the test mean squared error (MSE), a function of the variance plus the squared bias.
- Horizontal dash line: the minimum possible MSE.

More on ridge regression

- For $p \approx n$ and $p > n$,
  - OLS estimates are extremely variable
  - Ridge regression performs well by trading off a small increase in bias for a large decrease in variance.
- Computational advantages over best subset selection: for any fixed value of λ, ridge regression only fits a single model.

The Lasso

- One obvious disadvantage of Ridge regression:
  - The shrinkage penalty will never force any of the coefficient to be exactly zero.
  - The final model will include all variables, which makes it hard to interpret.
- The LASSO uses the $\mathcal{L}_1$ penalty to force some of the coefficient estimates to be exactly equal to zero, when the tuning parameter λ is sufficiently large.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_j\right)^2+\lambda\parallel\boldsymbol{\beta}\parallel_1=RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$

Ridge regression:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_j\right)^2+\lambda\parallel\boldsymbol{\beta}\parallel_2^2=RSS+\lambda\sum_{j=1}^{p}\beta_j^2$$

- The LASSO performs variable selection.

What does the $\mathcal{L}_1$ penalty do?

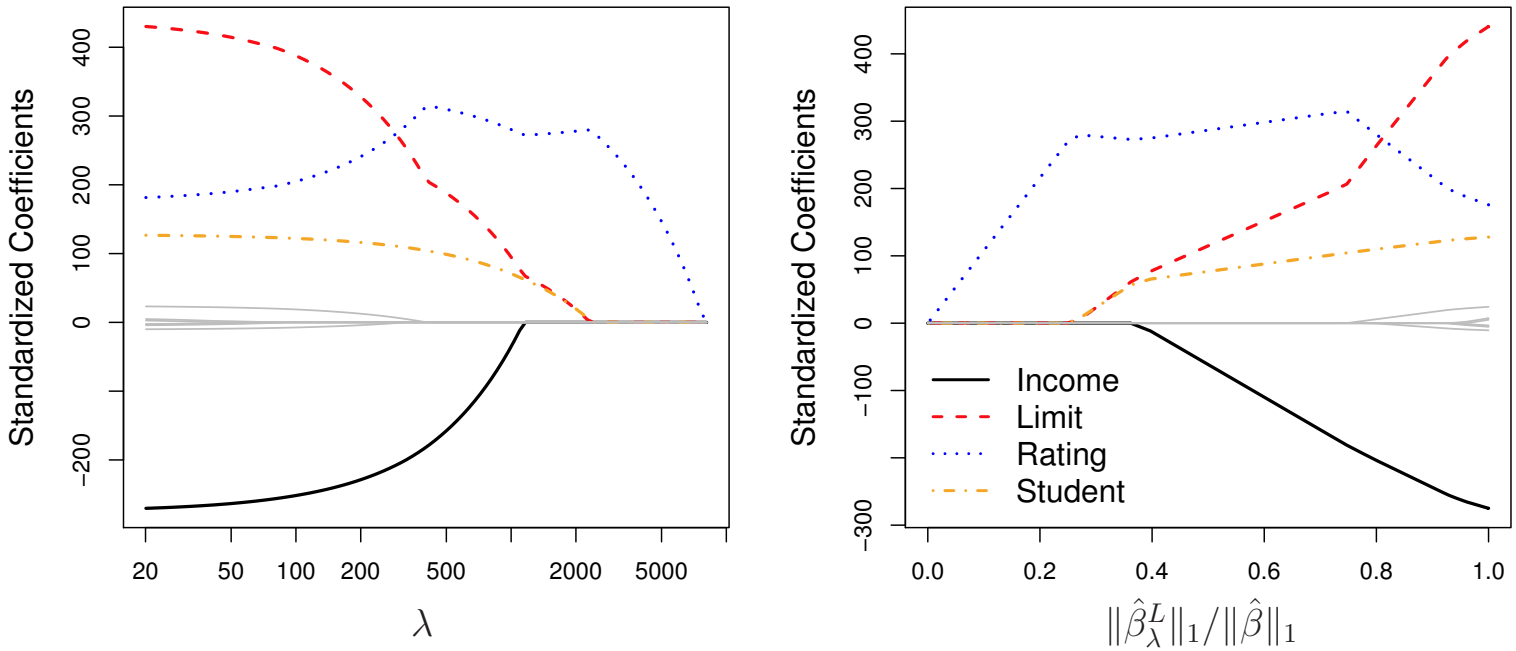


Figure: The standardized lasso coefficients on the Credit data set.

- When λ becomes sufficiently large, the lasso gives the null model.
- In the right-hand panel: rating $\Rightarrow$ rating + student + limit $\Rightarrow$ rating + student + limit + income

Comparing the Lasso and Ridge Regression

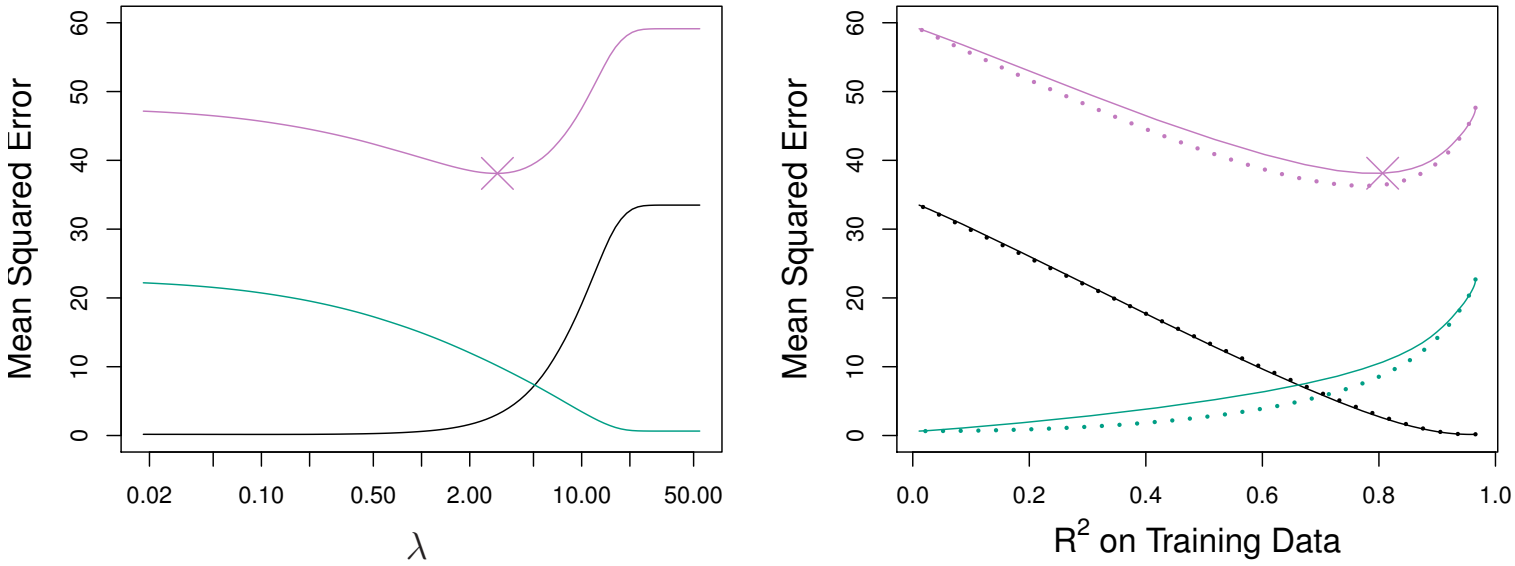


Figure: Plots of squared bias (black), variance (green), and test MSE (purple).

- Left plot: The lasso leads to qualitatively similar behavior to ridge regression.
- Right plot:
  - A plot against training R^2 can be used to compare models with different types of regularisation.
  - The minimum MSE of ridge regression is slightly smaller than that of the lasso
    - In the simulated dataset: all predicators were related to the response.

Comparing the Lasso and Ridge Regression

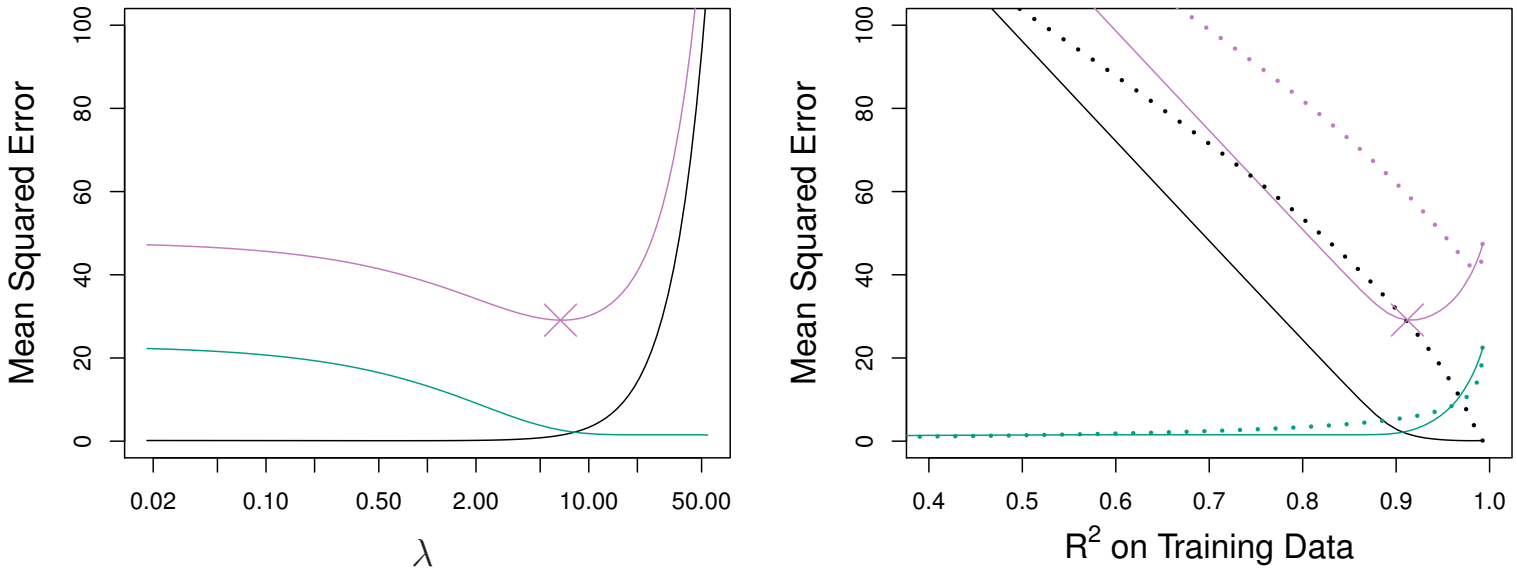


Figure: Plots of squared bias (black), variance (green), and test MSE (purple).

- In the above plots, the simulated data is generated in such a way that only 2 out of 45 predicators were related to the response.
- Conclusion: Neither ridge regression nor the lasso will universally dominate the other.

Conlcusions

- We expect:
  - The lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients.
  - Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size
- However, the number of predictors that is related to the response is never known a priori for real data sets.
  - Document classification: Given 10,000 words in a vocabulary, which words are related to document classes?
- Solution: cross validation!

Selecting the tuning parameter — 1

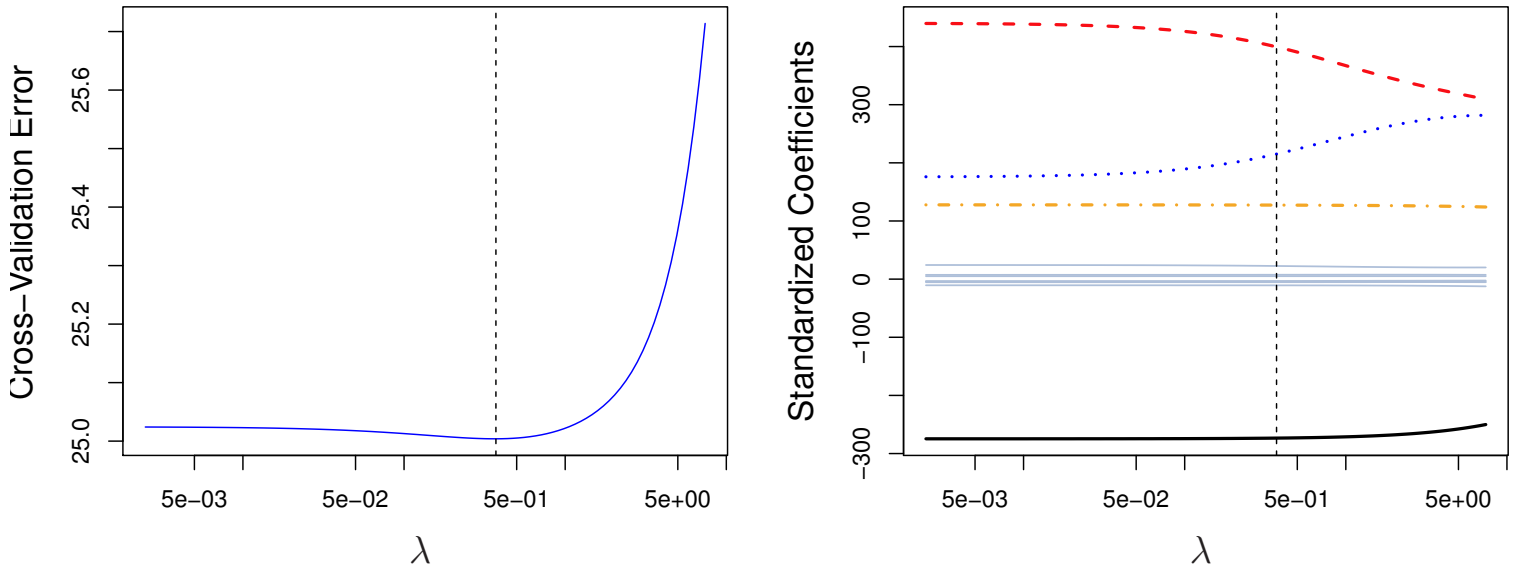


Figure: The choice of λ that results from performing leave- one-out cross-validation on the ridge regression fits from the Credit data set

- Select a grid of potential values, use cross validation to estimate the error rate on test data for each value of λ, and select the value that gives the least error rate
- Similar strategy applies to the Lasso.

Selecting the tuning parameter — 2

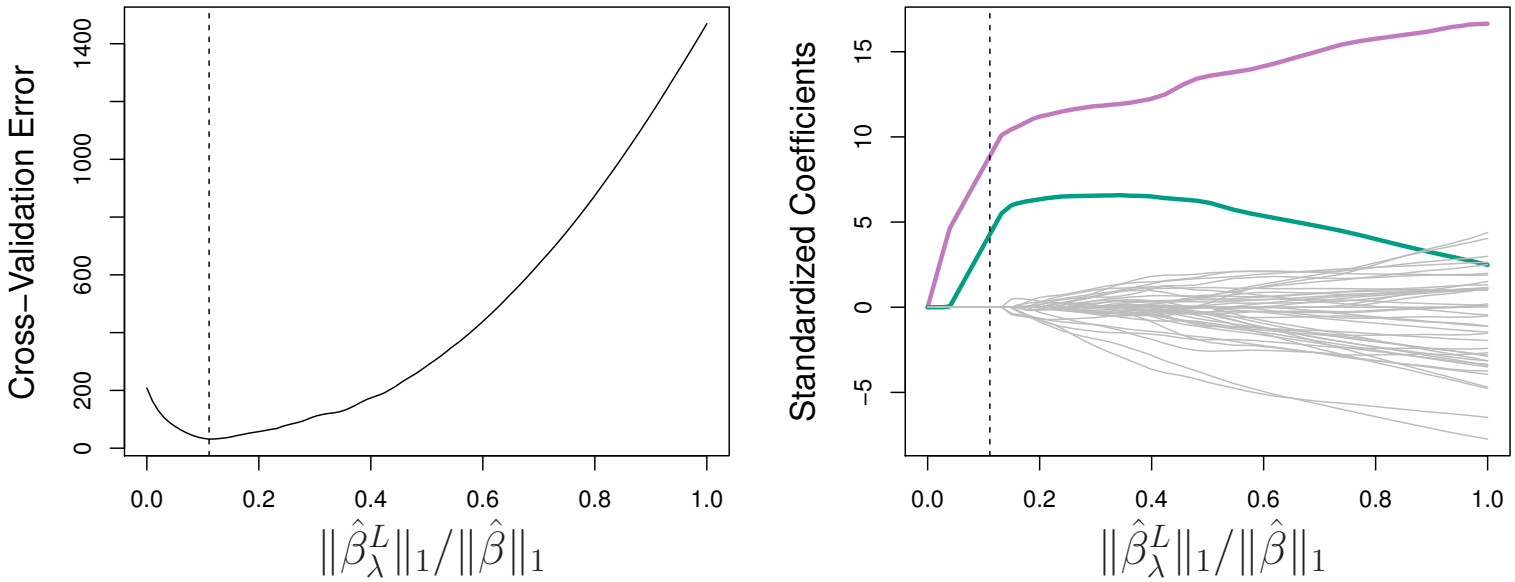


Figure: Left: Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set. Right: The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.

Outline

Elastic net

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda \parallel \boldsymbol{\beta} \parallel_1$$

- Limitations of the lasso
  - if $p > n$, the lasso selects at most n variables
  - Grouped variables: the lasso fails to do grouped selection.

Elastic net

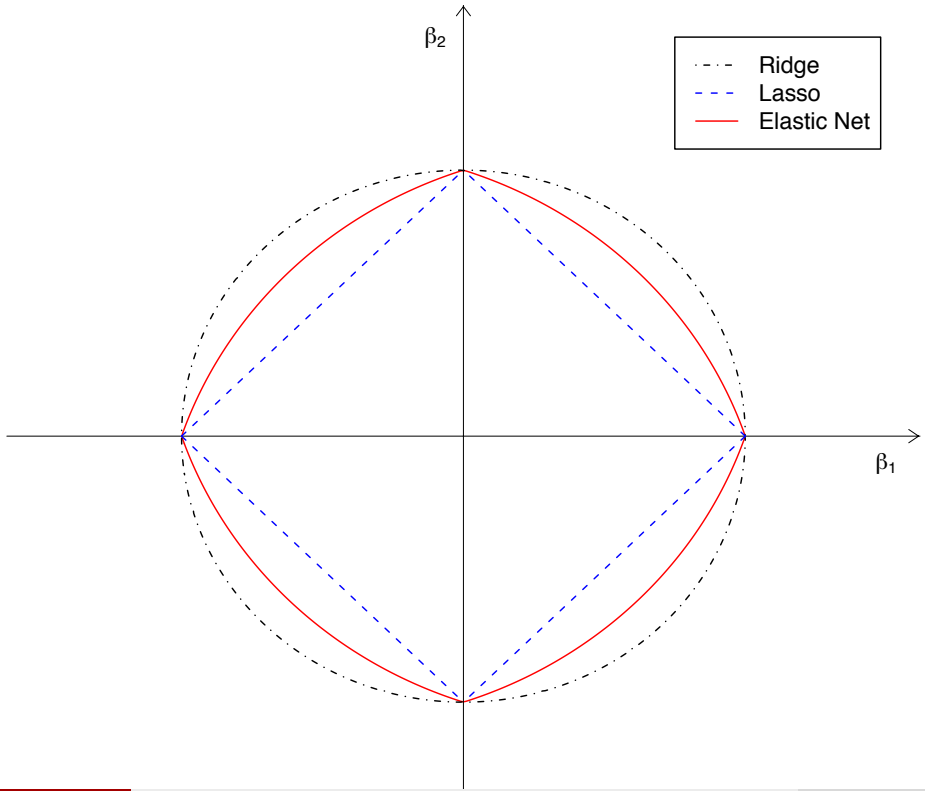
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\left((1-\alpha)\frac{\|\boldsymbol{\beta}\|_2^2}{2} + \alpha\|\boldsymbol{\beta}\|_1\right)$$

- The L_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty
  - Removes the limitation on the number of selected variables;
  - Encourages grouping effect;
  - Stabilizes the L_1 regularization path.
- Automatically include whole groups into the model if one variable amongst them is selected.

Elastic net

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\left((1-\alpha)\frac{\| \boldsymbol{\beta} \|_2^2}{2} + \alpha \| \boldsymbol{\beta} \|_1\right)$$

2-dimensional illustration $\alpha = 0.5$



Elastic net

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\left((1-\alpha)\frac{\| \boldsymbol{\beta} \|_2^2}{2} + \alpha \| \boldsymbol{\beta} \|_1\right)$$

- The elastic net performs simultaneous regularization and variable selection.
- Ability to perform grouped selection
- Appropriate for the $p >> n$ problem

Elastic net example

MONASH University

A simple illustration: elastic net vs. lasso

- Two independent "hidden" factors $\mathbf{z_1}$ and $\mathbf{z_2}$

$$\mathbf{z}_1 \sim U(0, 20), \quad \mathbf{z}_2 \sim U(0, 20)$$

- Generate the response vector $\mathbf{y} = \mathbf{z}_1 + 0.1 \cdot \mathbf{z}_2 + N(0, 1)$
- Suppose only observe predictors

$$\mathbf{x}_1 = \mathbf{z}_1 + \epsilon_1, \quad \mathbf{x}_2 = -\mathbf{z}_1 + \epsilon_2, \quad \mathbf{x}_3 = \mathbf{z}_1 + \epsilon_3$$
$$\mathbf{x}_4 = \mathbf{z}_2 + \epsilon_4, \quad \mathbf{x}_5 = -\mathbf{z}_2 + \epsilon_5, \quad \mathbf{x}_6 = \mathbf{z}_2 + \epsilon_6$$

- Fit the model on $(\mathbf{X}, \mathbf{y})$
- An "oracle" would identify $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$ (the $\mathbf{z}_1$ group) as the most important variables.

Figure: Slide from "Regularization and Variable Selection via the Elastic Net" by Zou and Hastie

Elastic net example

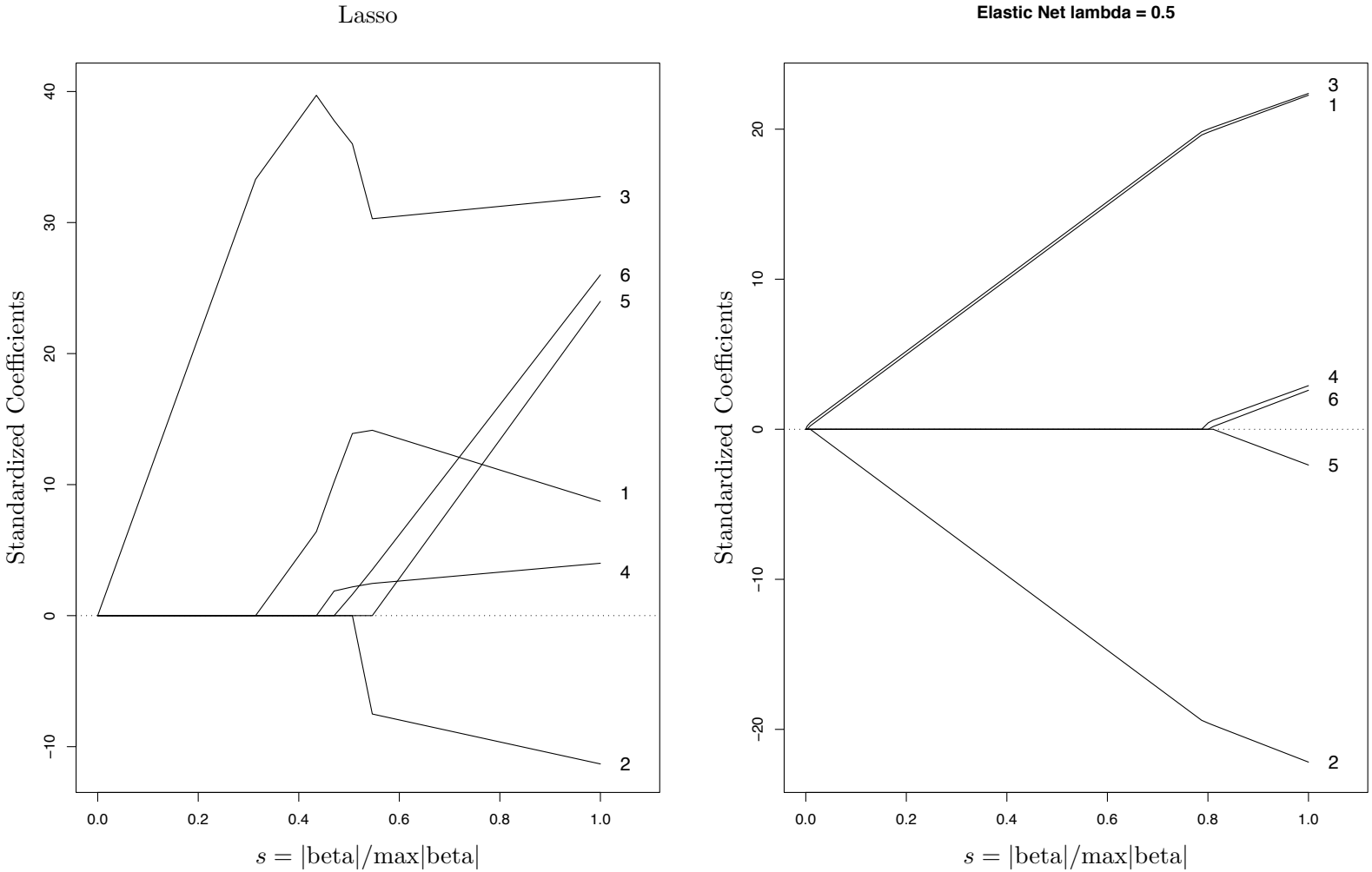


Figure: Slide from "Regularization and Variable Selection via the Elastic Net" by Zou and Hastie

Group Lasso

- Some advantages of group lasso
  - The information contained in the grouping structure is informative in learning.
  - Selecting important groups of variables gives models that are more sensible and interpretable
- Group Lasso formulation
  - We denote $\mathbf{X}$ as being composed of J groups $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_J$ with p_j denoting the size of group j; i.e., $\sum_j p_j = P$

$$\min_{\beta \in \mathbf{R}^P} \left(\| \mathbf{y} - \sum_j^L \mathbf{X}_j \boldsymbol{\beta}_j \|_2^2 + \lambda \sum_l^L \sqrt{p_l} \| \boldsymbol{\beta}_l \|_2 \right)$$

  - Group lasso acts like the lasso at the group level.
  - Group lasso does not yield sparsity within a group.

Sparse Group Lasso

- Sparse Group Lasso formulation
  - We denote $\mathbf{X}$ as being composed of J groups $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_J$ with p_j denoting the size of group j; i.e., $\sum_j p_j = P$

$$\min_{\beta \in \mathbf{R}^P} \left(\| \mathbf{y} - \sum_{j}^{L} \mathbf{X}_j \boldsymbol{\beta}_j \|_2^2 + \lambda_1 \sum_{l}^{L} \sqrt{p_l} \| \boldsymbol{\beta}_l \|_2 + \lambda_2 \| \boldsymbol{\beta} \|_1 \right)$$

  - Sparse Group lasso yields sparsity at both the group and individual feature levels.

Summary

- Model selection methods are an essential tool for data analysis, especially for big datasets involving many predictors.
- Reading materials:
  - "Linear Model Selection and Regularization", Chapter 6 of "Introduction to Statistical Learning", 6th edition
    - Section 6.1 "Subset Selection"
    - Section 6.2 "Shrinkage Methods"
- References:
  - Figures in this presentation were taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
  - Some of the slides are reproduced based on the slides from T. Hastie and R. Tibshirani