HKU BUSINESS SCHOOL
港大經管學院

# Business Statistics

# Model Selection and Regularization

**Zhanrui Cai**

Assistant Professor in Analytics and Innovation

ISLR Chapter 5, 6

# Example – Credit Data Set (Page 83)

- Balance: average credit card debt for a number of individuals

- Age, Gender, Ethnicity (Caucasian, African American or Asian)
- Cards: number of credit cards
- Education: years of education
- Income (in thousands of dollars)
- Limit: credit limit
- Rating: credit rating
- Student (student status), Status (marital status)

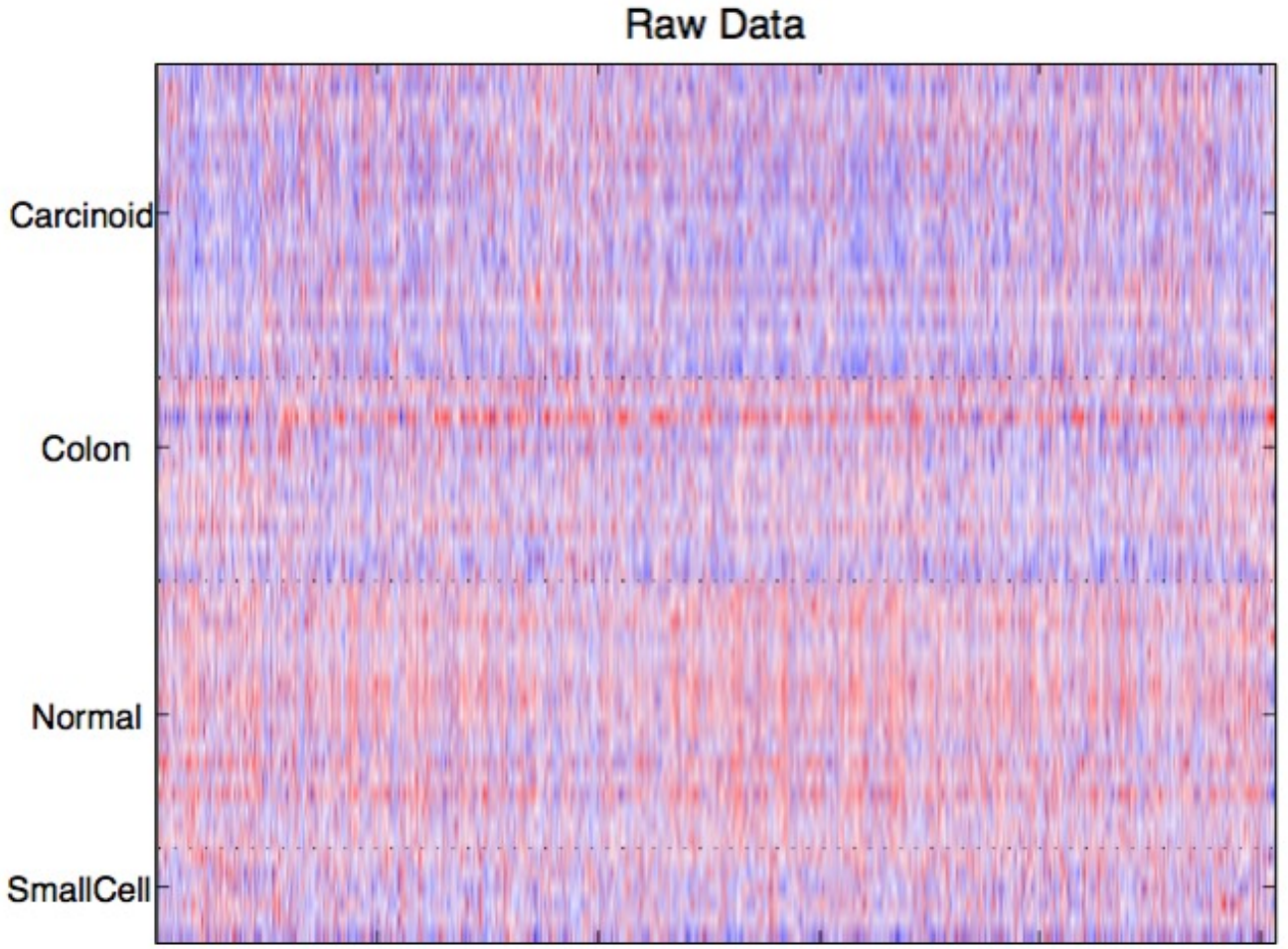
# Lung Cancer Gene Expression Microarray

- Genes: 12,625
- Samples: 56

- 4 Subgroups
  - Normal
  - Pulmonary Carcinoid Cancer
  - Colon Cancer
  - Small Cell Cancer

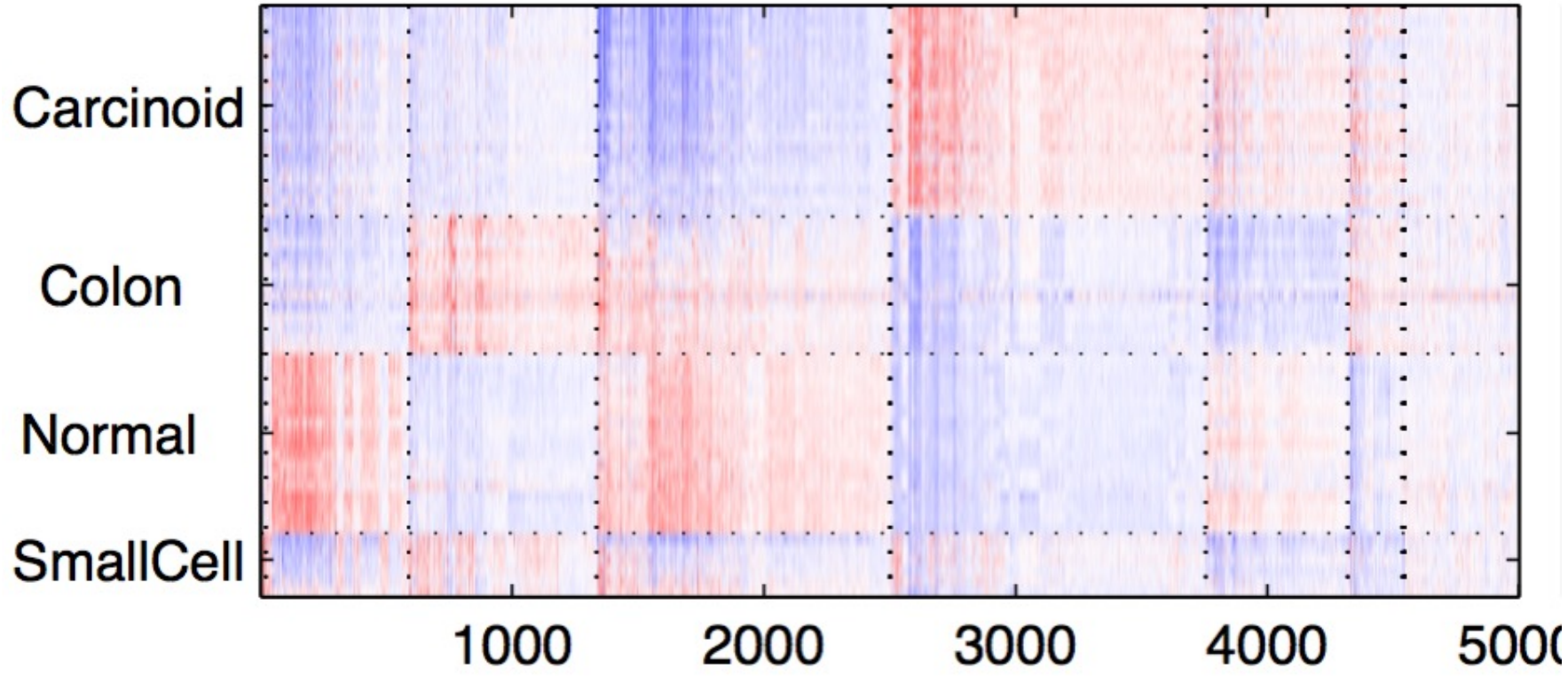
- What genes can be used to predict cancer type?

# Lung Cancer Gene Expression Microarray, Heatmap




- No clear association!

# Lung Cancer Microarray, Gene Selection




- Select 5,000 relevant genes, others “irrelevant”
- Reorder the genes, reveal clear checkboard (i.e. association)

# Linear Model Selection

- Recall the linear model

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

- Model selection – select the relevant $X$ features.
  - Number of $X$ variables
  - Which $X$ variables

- Prediction Accuracy: especially when $p > n$, to control the variance and enable model fitting.

- Model Interpretability: by removing irrelevant features through setting the corresponding coefficients to be zero.

# Normal Mean Problem: Least Squares

- Consider $y_1, \ldots, y_n \sim Normal(\mu, \sigma^2)$, i.i.d.
- Estimation of $\mu, \sigma^2$
- Formulate this as linear reg with just intercept

$$y_i = \mu + \varepsilon_i, \quad \varepsilon_i \; i.i.d. N(0, \sigma^2)$$

- Least squares estimation

$$\hat{\mu}_{LS} = argmin_{\mu} \sum_{i=1}^{n} (y_i - \mu)^2 = \bar{y}$$

$$\hat{\sigma}^2{}_{LS} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}$$

# Normal Mean Problem: Maximum Likelihood

- Note the normal density function:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

  which is the likelihood function of $\mu, \sigma^2$ given data $y_i$

- Given independence, the joint likelihood is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}}$$

- Maximum likelihood estimation

$$\hat{\mu}_{ML} = argmax_{\mu} L(\mu, \sigma^2) = \bar{y}$$

$$\hat{\sigma}^2{}_{ML} = argmax_{\sigma^2} L(\mu, \sigma^2) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

# Least Squares and Maximum Likelihood

- For normal errors, the least squares estimate and the maximum likelihood estimate for $\mu$ are the same.

- For normal errors, the least squares estimate and the maximum likelihood estimate for $\sigma^2$ differ only in the denominator: $n-1$ and $n$.

- The same hold for linear reg with $p$ $X$ variables.
  - Same estimate for the coefficients $\beta_j$
  - $\hat{\sigma}^2{}_{LS} = \frac{RSS}{n-p-1}$
  - $\hat{\sigma}^2{}_{ML} = \frac{RSS}{n}$

# Three Classes of Model Selection Methods

- *Subset Selection*. We identify a subset of the $p$ predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.
- *Shrinkage*. We fit a model involving all $p$ predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as *regularization*) has the effect of reducing variance and can also perform variable selection.
- *Dimension Reduction*. We project the $p$ predictors into a $M$-dimensional subspace, where $M < p$. This is achieved by computing $M$ different *linear combinations*, or *projections*, of the variables. Then these $M$ projections are used as predictors to fit a linear regression model by least squares.

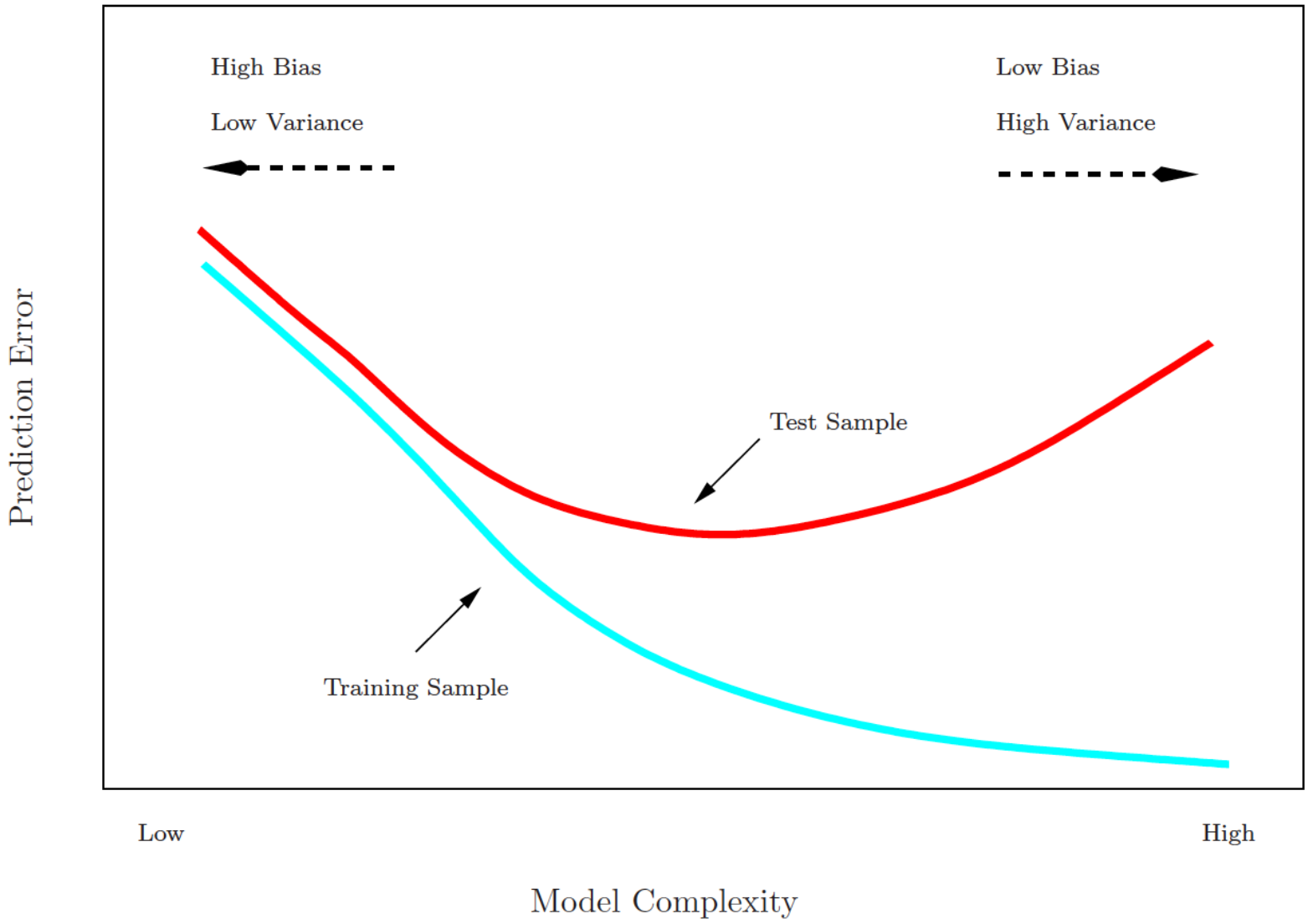
# Choosing the Optimal Model

- The model containing all of the predictors will always have the smallest RSS and the largest $R^2$, since these quantities are related to the training error.

- We wish to choose a model with low test error, not a model with low training error. Recall that training error is usually a poor estimate of test error.

- Therefore, RSS and $R^2$ are not suitable for selecting the best model among a collection of models with different number of predictors.

# Model Complexity & Prediction Error



# Estimating Test Error: Two Approaches

1. Indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.

2. Directly estimate the test error, using either a validation set approach or a cross-validation approach.

# Model Comparison Criteria

## Adjustment to the Training Error

- Adjusted RSS
- Mallow's $C_p$
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

# Adjusted $R^2$

- For a least squares model with $p$ variables, the Adjusted $R^2$ statistic is calculated as

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$$

  where TSS is the total sum of squares.

- A large Adjusted $R^2$ indicates a model with a small test error.

- Maximizing the Adjusted $R^2$ is equivalent to minimizing $\frac{\text{RSS}}{n-p-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{\text{RSS}}{n-p-1}$ may increase or decrease.

- Unlike $R^2$, Adjusted $R^2$ pays a price for the inclusion of unnecessary variables in the model.

# Mallow's $C_p$

- Let $\hat{\sigma}^2$ denote an estimate of the variance of the error $\epsilon$ obtained from the fitted model.
- $p$ : # of predictors in the model.
- Mallow's $C_p$:

$$C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2).$$

- Select the model with the smallest $C_p$.

# Bayesian Information Criterion

- The BIC is defined as

$$BIC = \frac{1}{n}(\text{RSS} + \log(n)p\hat{\sigma}^2).$$

- Select the model with the lowest BIC value.

- Notice that BIC replaces the $2p\hat{\sigma}^2$ used by $C_p$ with a $\log(n)p\hat{\sigma}^2$ term, where $n$ is the number of observations.

- Since $\log(n) > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than $C_p$.

# Akaike Information Criterion

- The AIC is defined for a large class of models fit by maximum likelihood:

$$AIC = -2 \log L + 2 \cdot p$$

where $L$ is the maximized value of the likelihood of the estimated model.

- Select the model with the smallest AIC

- For linear models with Gaussian errors
  - maximum likelihood = least squares
  - $C_p$ and AIC: equivalent

# Subset Selection

*Best subset and stepwise model selection procedures*

*Best Subset Selection*

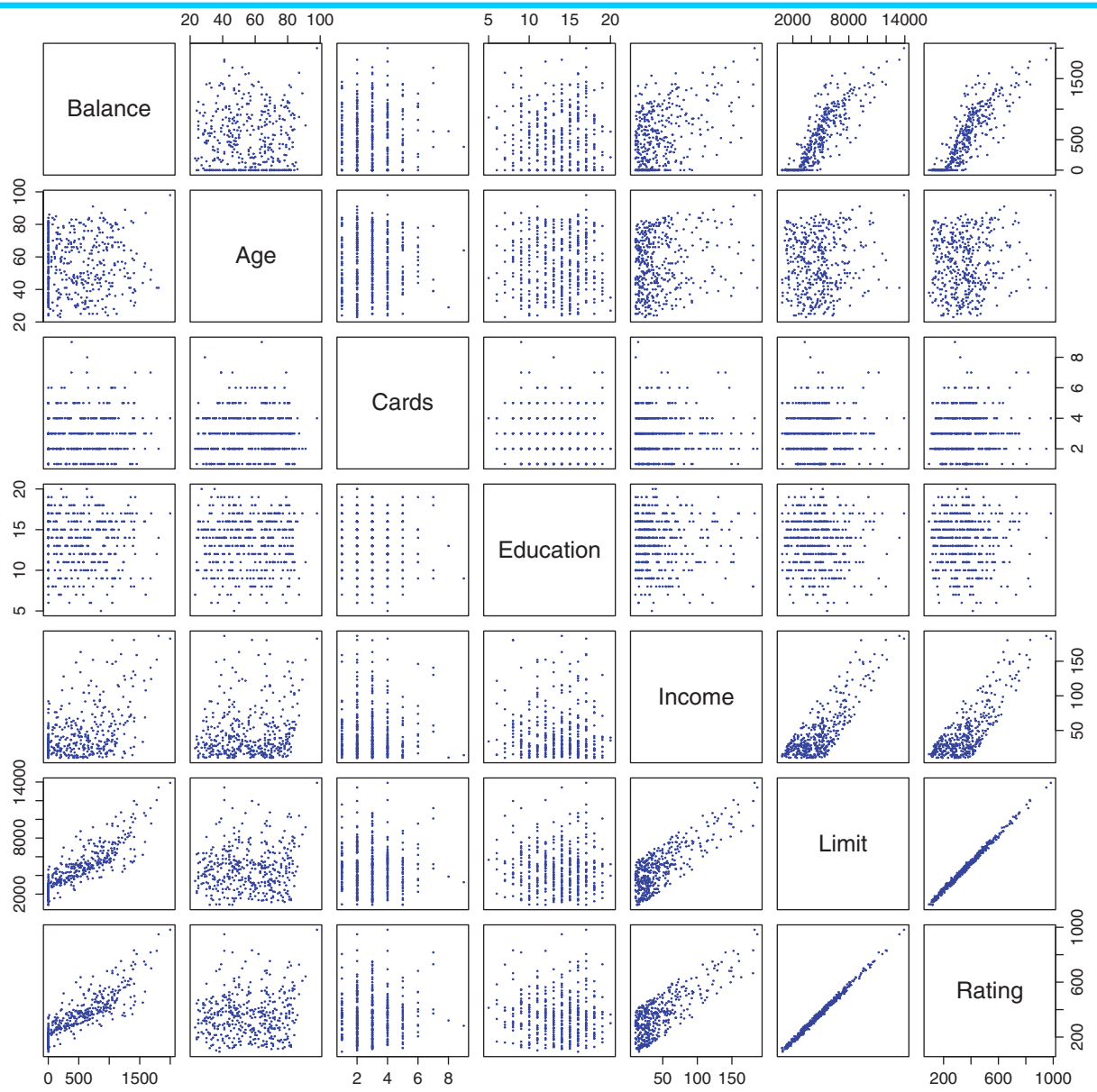
1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, 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 (Page 83)

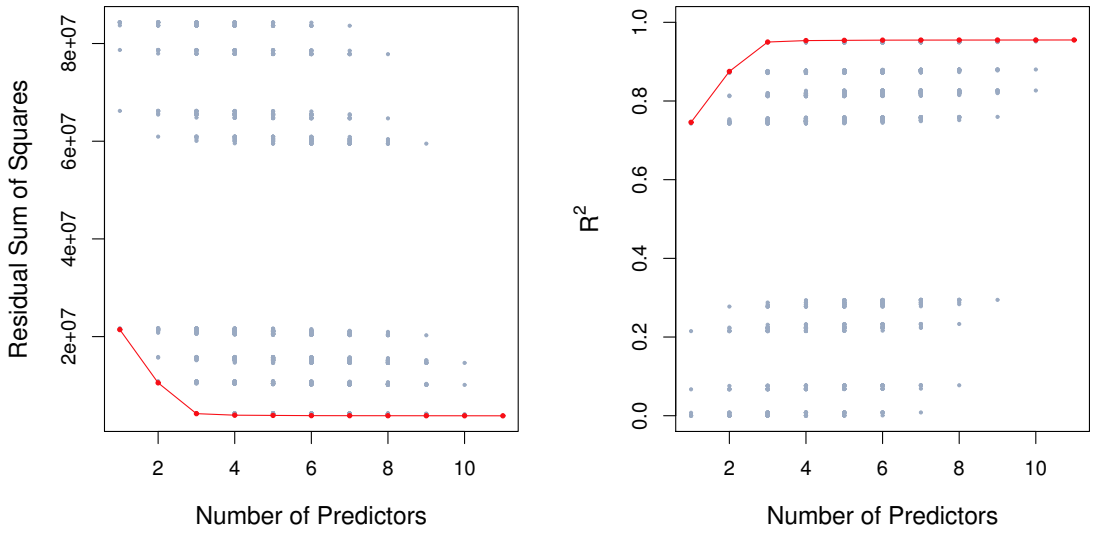
- Balance: average credit card debt for a number of individuals

- Age, Gender, Ethnicity (Caucasian, African American or Asian)
- Cards: number of credit cards
- Education: years of education
- Income (in thousands of dollars)
- Limit: credit limit
- Rating: credit rating
- Student (student status), Status (marital status), Own (house ownership), region (East, West or South)

# Scatter Plot Matrix

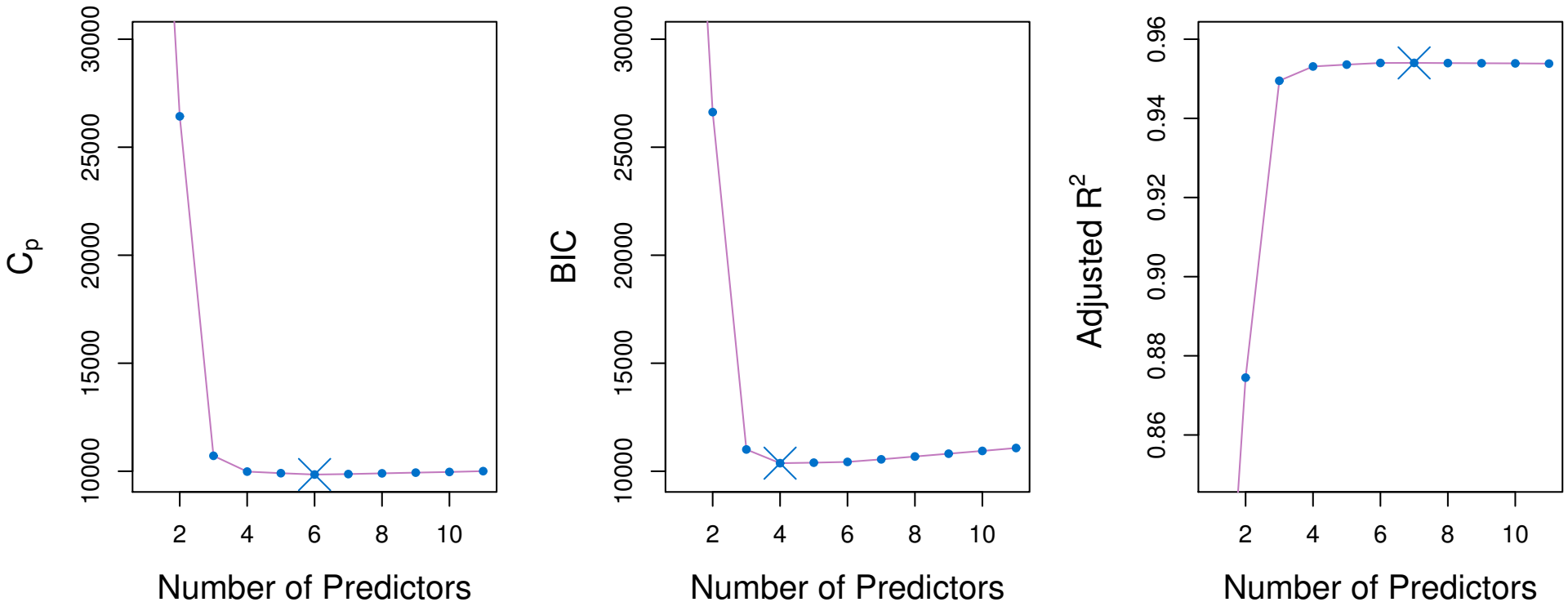


# Example – Credit Data Set




*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*

# Credit Data Example: Best Subset



# Extension to Other Models

- Although we have presented best subset selection here for least squares regression, the same ideas apply to other types of models, such as logistic regression (to be discussed).

- The deviance — negative two times the maximized log-likelihood — plays the role of RSS for a broader class of models.

# Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large *p*.
- Best subset selection may also suffer from statistical problems when *p* is large:
  - The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- An enormous search space can lead to overfitting and high variance of the coefficient estimates.
- For both reasons, stepwise methods, which explore a far more restricted set of models, are attractive alternatives.

# Forward Stepwise Selection

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
- In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.
- Computational advantage over best subset selection is clear.
- It is not guaranteed to find the best possible model out of all models containing subsets of the $p$ predictors.

# Forward Stepwise Selection: Details

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   - 2.1 Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   - 2.2 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 $R^2$.

# Example: Credit Data (Page 231)

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

*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

- Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection.

- However, unlike forward stepwise selection, it begins with the full least squares model containing all $p$ predictors, and then iteratively removes the least useful predictor, one-at-a-time.

# Backward Stepwise Selection: Details

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all $p$ predictors.
2. For $k = p, p-1, \ldots, 1$:
   - 2.1 Consider all $k$ models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   - 2.2 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

- 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). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when $p$ is very large.

# Extensions

- Hybrid-stepwise selection considers both forward and backward moves at each step, and selects the best of the two.
  - Pros: computationally efficient, error made at an earlier stage can be corrected later.
  - Need a criterion to decide whether to add or drop at each step. e.g. AIC takes proper account of both the number of parameters and how good the model fits.

# Validation Set and Cross-Validation

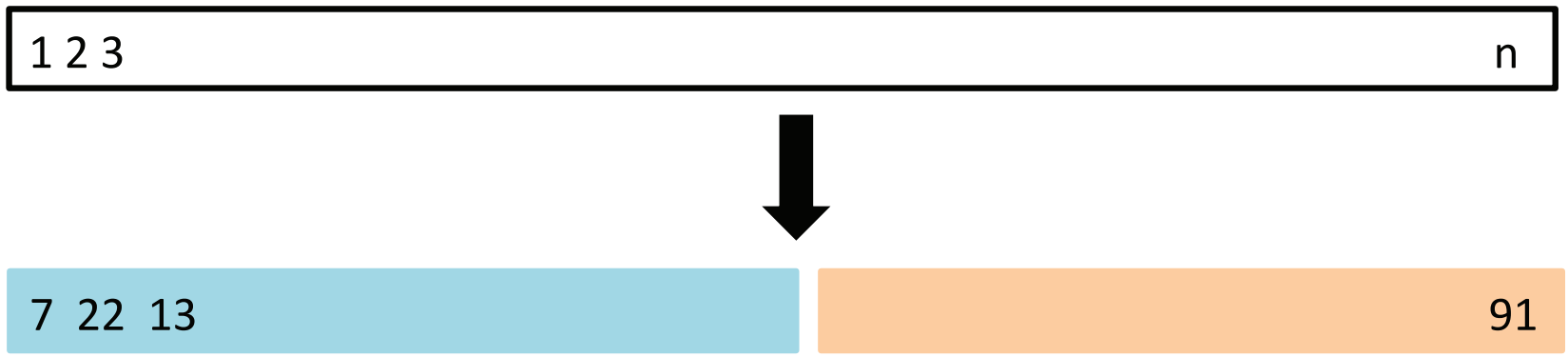
# Validation and Cross-Validation

- Each of the procedures returns a sequence of models $\mathcal{M}_k$ indexed by model size $k = 0, 1, 2, \ldots$. Our job here is to select $\hat{k}$. Once selected, we will return model $\mathcal{M}_{\hat{k}}$
- We compute the validation set error or the cross-validation error for each model $\mathcal{M}_k$ under consideration, and then select the $k$ for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to AIC, BIC, $C_p$, and adjusted $R^2$, in that it provides a direct estimate of the test error, and *doesn't require an estimate of the error variance* $\sigma^2$.
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance $\sigma^2$.

# Validation-set Approach

- We would like to pick the model that has smallest testing error. But no test data is available at the training stage!
- Create an artificial testing data from training data!
- We randomly divide the available set of samples into two parts: a training set and a validation or hold-out set.
- The model is fit on the training set, and then used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error. Use MSE.

# The Validation Process




A random splitting into two halves: left part is training set, right part is validation set

# Drawbacks of the Validation Set Approach

- Only a subset of the observations are used to fit the model.
- The validation estimate of the test error can be highly variable, depending on the split of the raw data.
- The validation test error may tend to overestimate the test error for the model fit on the entire data set.
- Cross-validation (CV) to the rescue!

# K-fold Cross-Validation (CV)

- *Widely used approach* for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into $K$ equal-sized parts. We leave out part $k$, fit the model to the other $K-1$ parts (combined), and then obtain predictions for the left-out $k$th part.
- This is done in turn for each part $k = 1, 2, \ldots K$, and then the results are combined.

# K-fold CV Illustration

Divide data into $K$ roughly equal-sized parts ($K = 5$ here)

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Validation | Train | Train | Train | Train |

# The Computing Details

- Let the $K$ parts be $C_1, C_2, \ldots C_K$, where $C_k$ denotes the indices of the observations in part $k$. There are $n_k$ observations in part $k$: if $N$ is a multiple of $K$, then $n_k = n/K$.
- Compute

$$\text{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k$$

  where $\text{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and $\hat{y}_i$ is the fit for observation $i$, obtained from the data with part $k$ removed.
- Setting $K = n$ yields $n$-fold or *leave-one out cross-validation* (LOOCV).

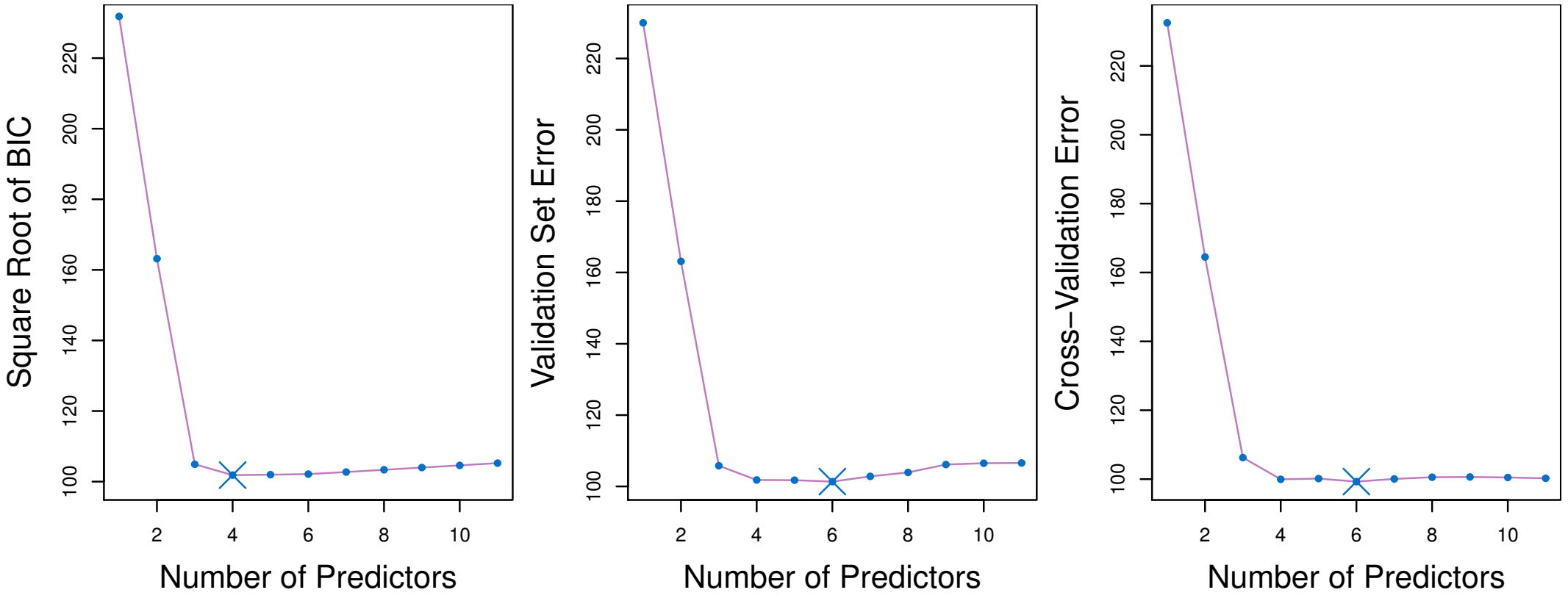
# A Nice Special Case!

With least-squares linear or polynomial regression, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit! The following formula holds:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

where $\hat{y}_i$ is the $i$th fitted value from the original least squares fit, and $h_i$ is the leverage (diagonal of the "hat" matrix; see book for details.) This is like the ordinary MSE, except the $i$th residual is divided by $1 - h_i$.

# Credit Data Example: Validation



# Details of Previous Figure

- The validation errors were calculated by randomly selecting three-quarters of the observations as the training set, and the remainder as the validation set.
- The cross-validation errors were computed using $k = 10$ folds.
- In this case, the validation and cross-validation methods both result in a six-variable model.
- However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

# Shrinkage Methods

# Shrinkage Methods

- 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.
- Ridge regression and Lasso regression
- 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

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

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

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

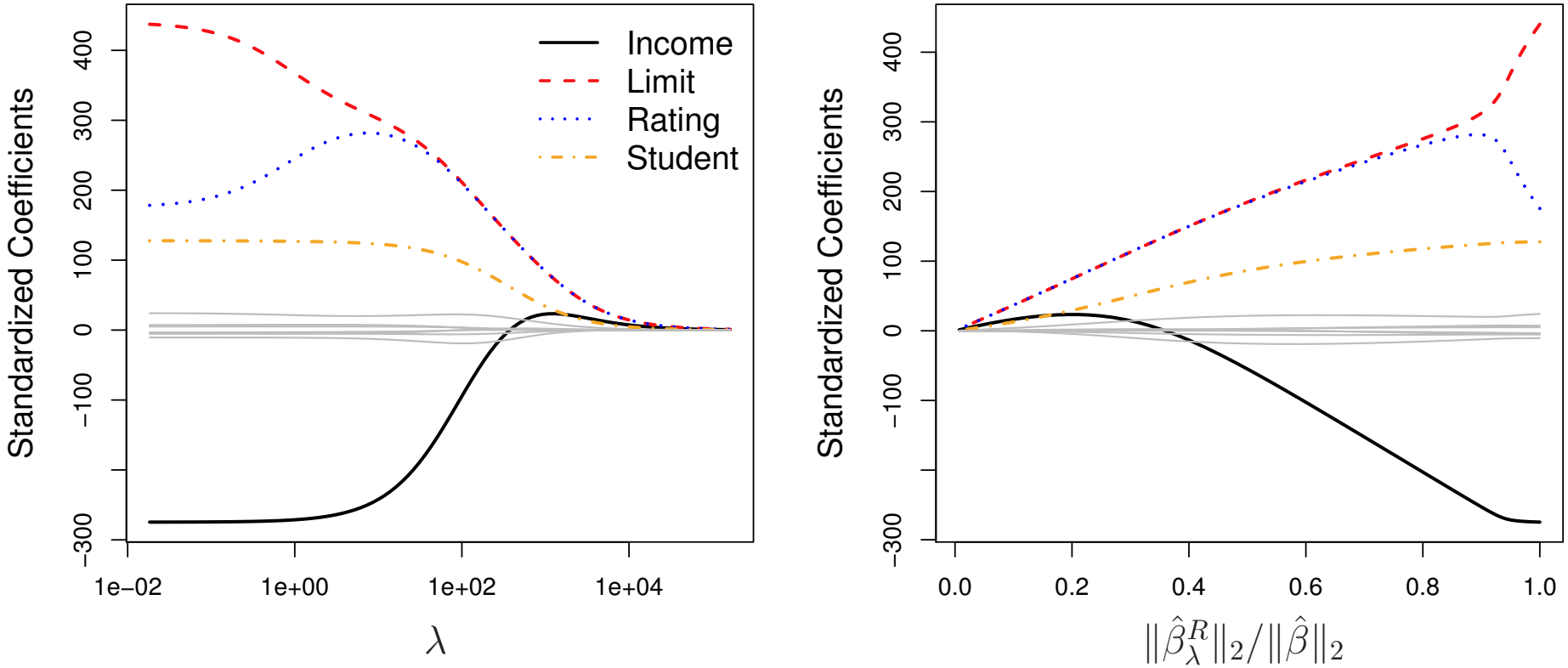
$$\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately.

# Ridge Regression Continued

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
- However, the second term, $\lambda \sum_j \beta_j^2$, called a *shrinkage penalty*, is small when $\beta_1, \ldots, \beta_p$ are close to zero, and so it has the effect of *shrinking* the estimates of $\beta_j$ towards zero.
- The tuning parameter $\lambda$ serves to control the relative impact of these two terms on the regression coefficient estimates.
- Selecting a good value for $\lambda$ is critical; cross-validation is used for this.

# Example: Credit Data




Grey lines: unrelated variables, i.e. noise variables, in contrast with signal variables (colored ones).

# Details of Previous Figure

- In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of $\lambda$.
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but instead of displaying $\lambda$ on the $x$-axis, we now display $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$, where $\hat{\beta}$ denotes the vector of least squares coefficient estimates.
- The notation $\|\beta\|_2$ denotes the $\ell_2$ norm (pronounced "ell 2") of a vector, and is defined as $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} {\beta_j}^2}$.

# Ridge Regression: Scaling of Predictors

- The standard least squares coefficient estimates are *scale equivariant*: multiplying $X_j$ by a constant $c$ simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. In other words, regardless of how the $j$th predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can change *substantially* when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after *standardizing the predictors*, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

# The Bias-Variance Decomposition

- Assume that

$$Y = f(X) + \varepsilon$$

where E($\varepsilon$)=0 and Var($\varepsilon$)=$\sigma_\varepsilon^2$.

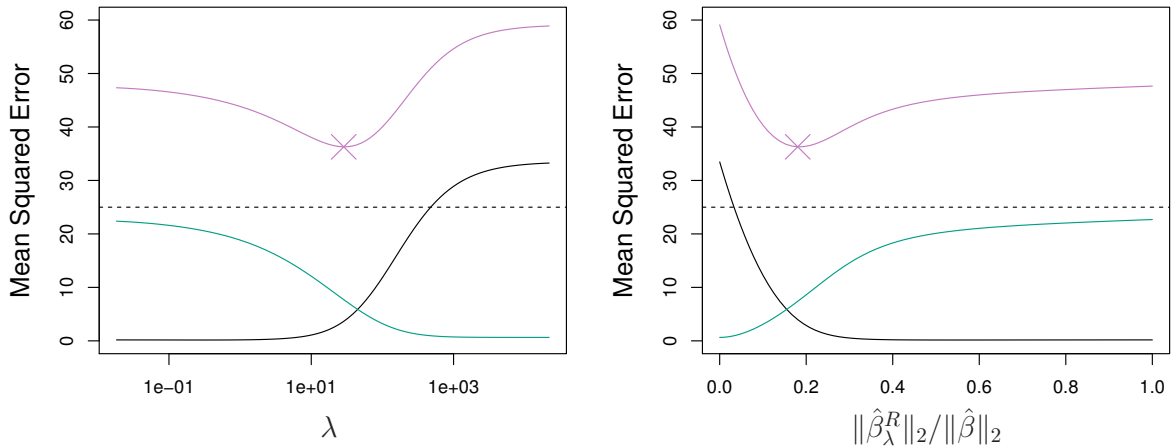
- At an input point $X = x_0$, the expected squared prediction error is

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \mathrm{Bias}^2 + \text{Variance}.
\end{aligned}$$

- The more complex the model, the lower the bias but the higher the variance.

# Ridge Regression Improves Over Least Squares

*The Bias-Variance tradeoff*




*Simulated data with $n = 50$ observations, $p = 45$ predictors, all having nonzero coefficients. 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 $\lambda$ 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.*

# The Lasso

- Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all $p$ predictors in the final model
- The *Lasso* is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

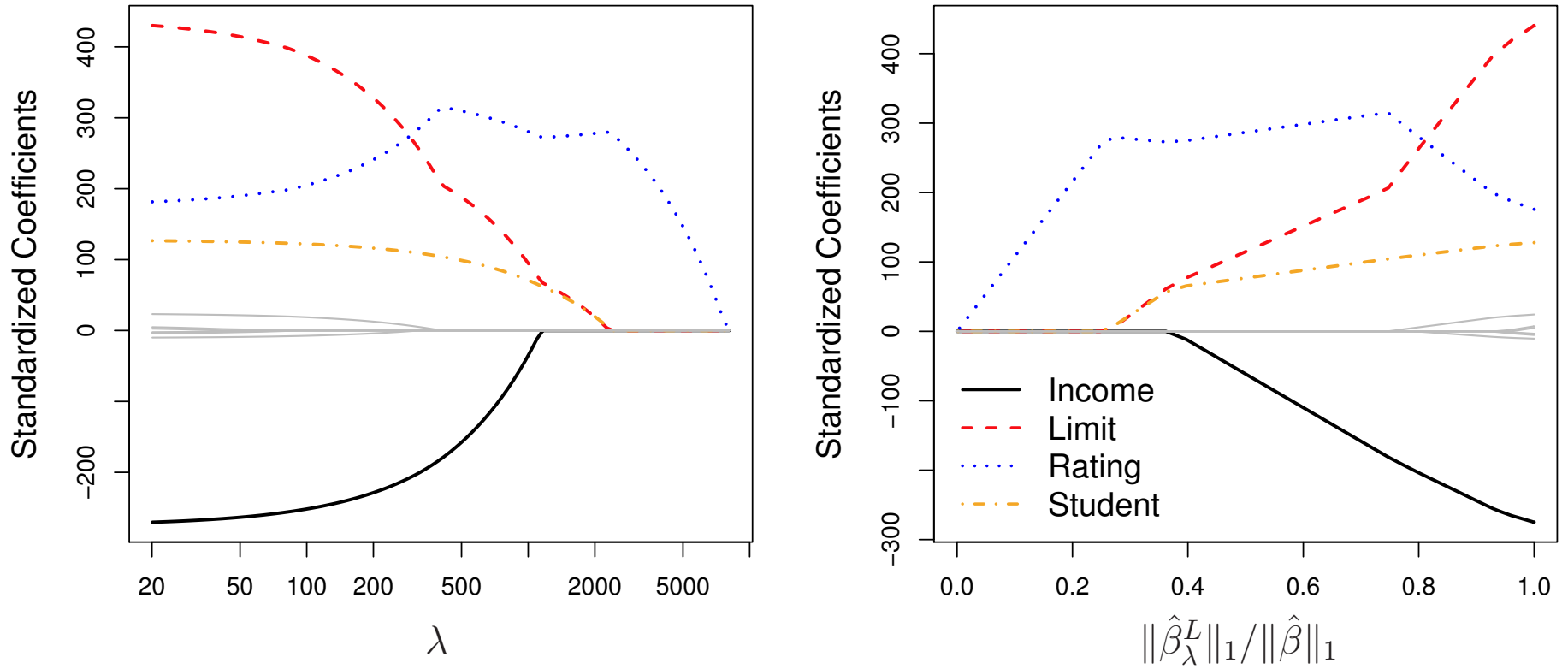
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

- In statistical parlance, the lasso uses an $\ell_1$ (pronounced "ell 1") penalty instead of an $\ell_2$ penalty. The $\ell_1$ norm of a coefficient vector $\beta$ is given by $\|\beta\|_1 = \sum|\beta_j|$.

# The Lasso Continued

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the $\ell_1$ penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter $\lambda$ is sufficiently large.
- Hence, much like best subset selection, the lasso performs *variable selection*.
- We say that the lasso yields *sparse* models — that is, models that involve only a subset of the variables.
- As in ridge regression, selecting a good value of $\lambda$ for the lasso is critical; cross-validation is again the method of choice.

# Example Credit Data



# The Variable Selection Property of the Lasso

Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?

One can show that the lasso and ridge regression coefficient estimates solve the problems

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

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

respectively.

# Best Subset, Lasso, and Ridge

- Best Subset

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s.$$

- Lasso

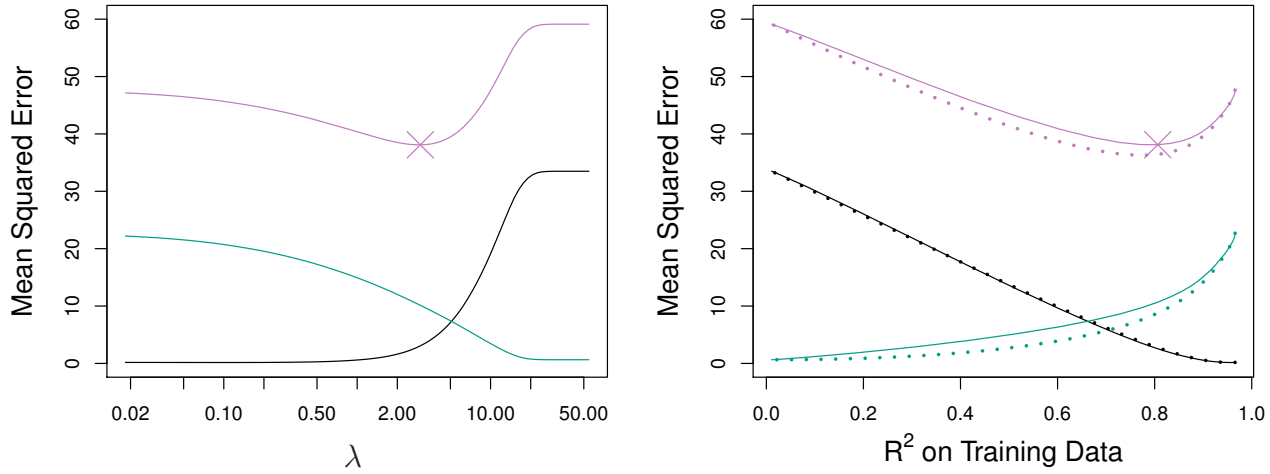
$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

- Ridge

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

# Lasso vs Ridge Regression

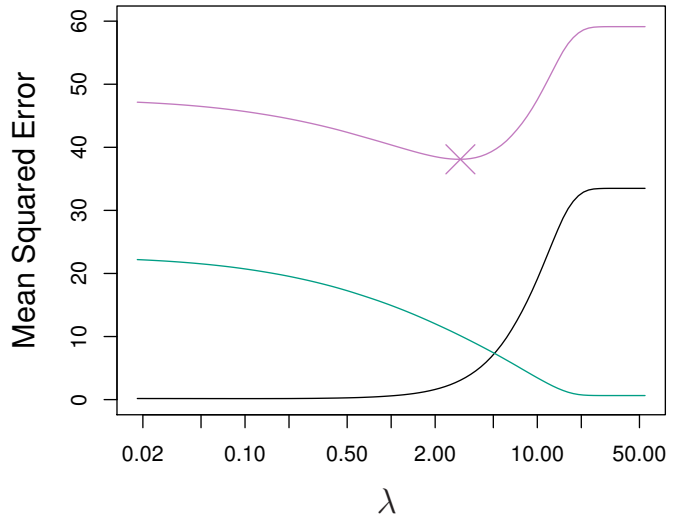
## Simulation 1: n=50, p=45, all relevant

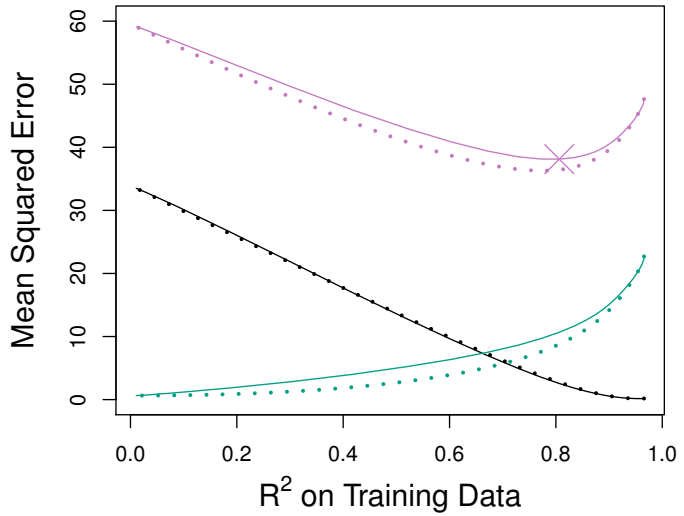



*Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on simulated data set of Slide 32. Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). Both are plotted against their $R^2$ on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

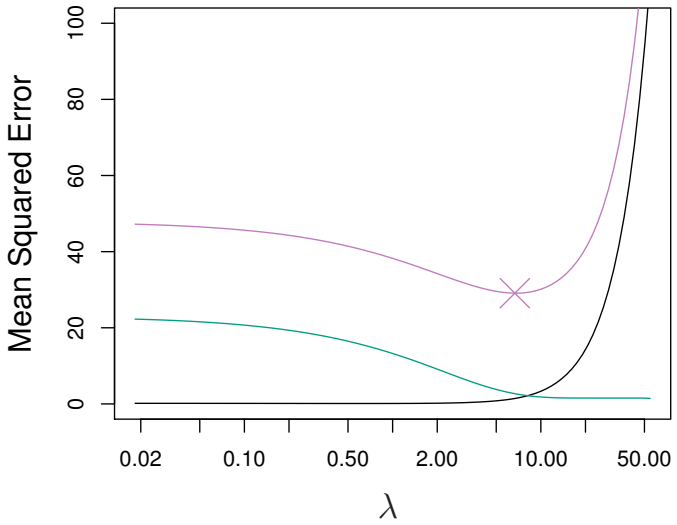
# Lasso vs Ridge Continued

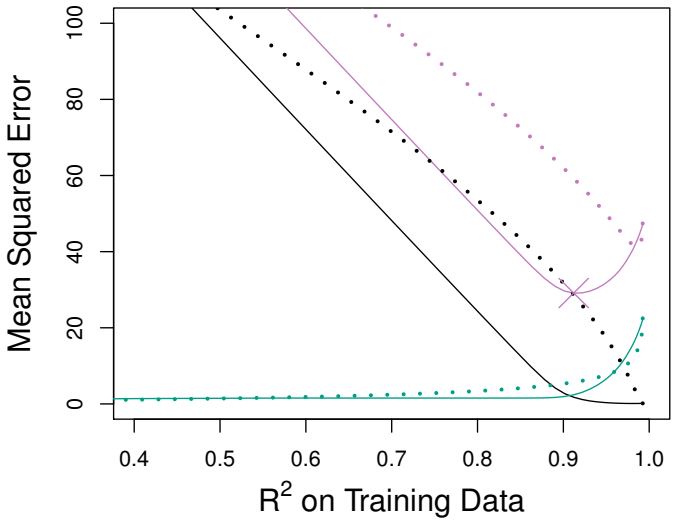
## Simulation 1: n=50, p=45, all relevant




## Simulation 2: n=50, p=45, only 2 relevant



# Comparison Conclusions

- These two examples illustrate that neither ridge nor the lasso will universally dominate the other.
- In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors.
- However, the number of predictors that is related to the response is never known a priori for real data sets.
- A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.

# A Simple Special Case

- Consider $n = p, X = I$ with no intercept.
- The least squares regression minimizes $\sum_{j=1}^{p}(y_j - \beta_j)^2$ with solution $\widehat{\beta_j} = y_j$.
- The ridge regression minimizes $\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$ with solution

$$\hat{\beta}_j^R = y_j/(1+\lambda).$$

- The lasso regression minimizes $\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$ with solution

$$\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} & if \quad y_j > \frac{\lambda}{2}; \\ y_j + \frac{\lambda}{2} & if \quad y_j < -\frac{\lambda}{2} \\ 0 & if \quad |y_j| \leq \frac{\lambda}{2}. \end{cases}$$

# Geometry of Ridge and Lasso

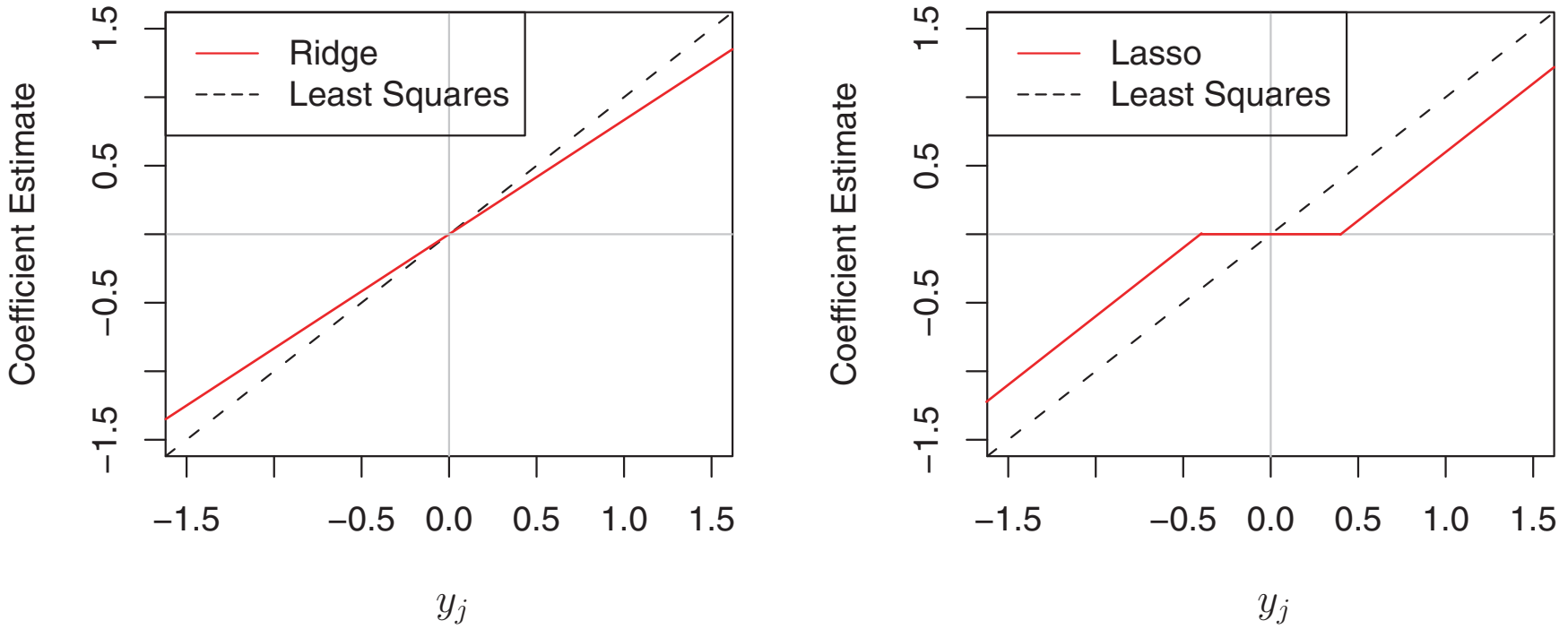
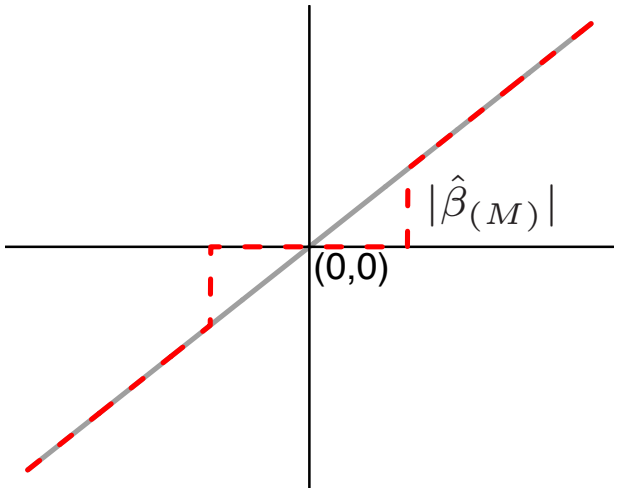



**FIGURE 6.10.** *The ridge regression and lasso coefficient estimates for a simple setting with* $n = p$ *and* $\mathbf{X}$ *a diagonal matrix with* 1*'s on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: *The lasso coefficient estimates are soft-thresholded towards zero.*

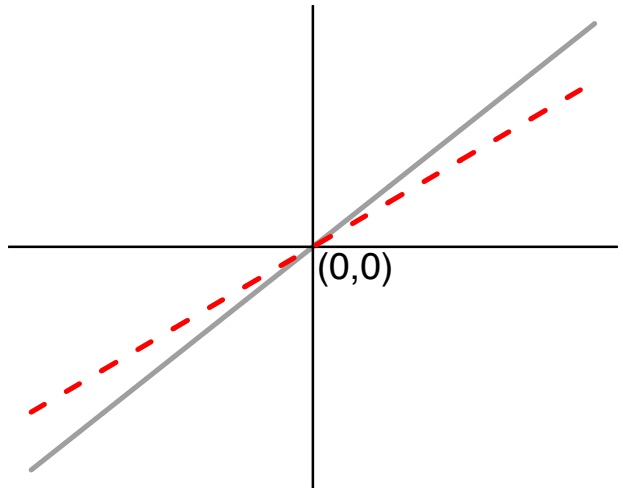
# Cases with Orthonormal Design Matrix $X$

| 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 | $\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$ |

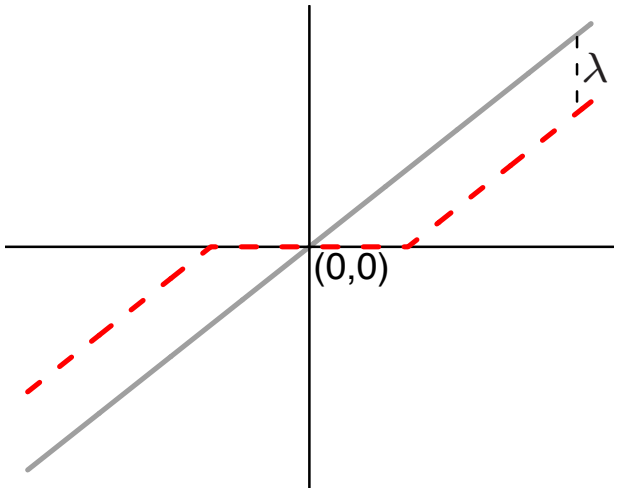
Best Subset




Ridge



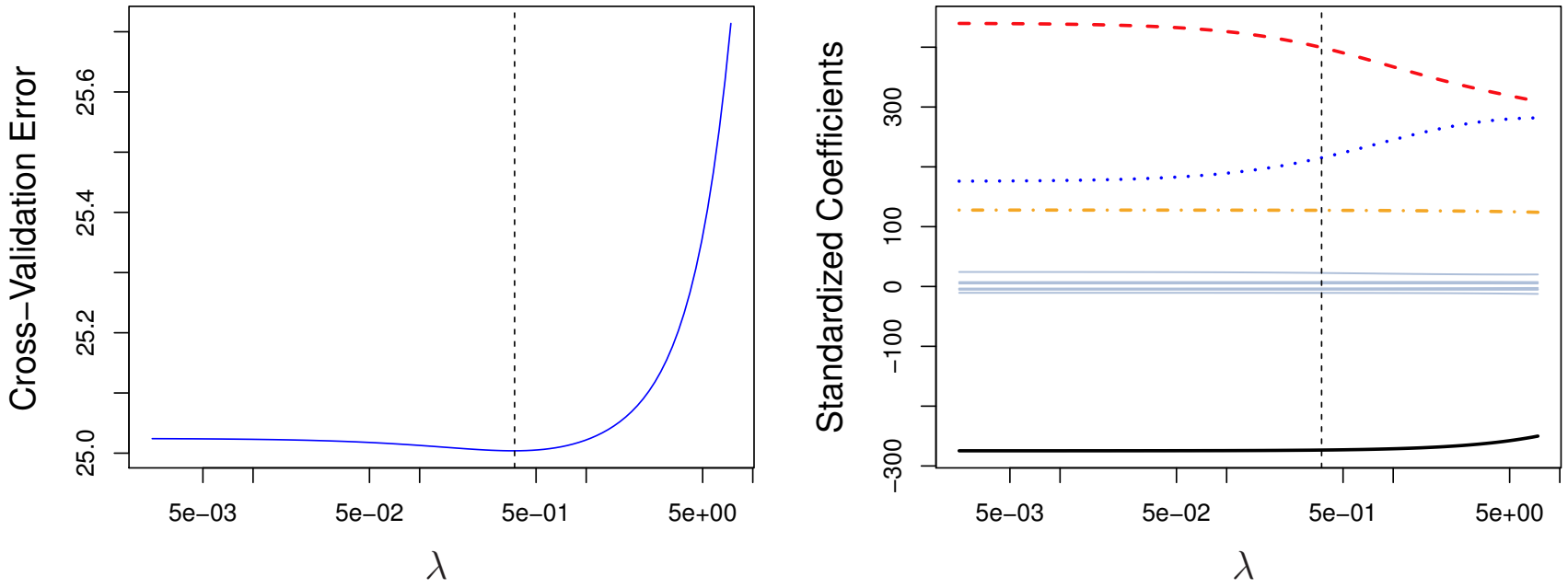

Lasso



# Tuning Parameter Selection

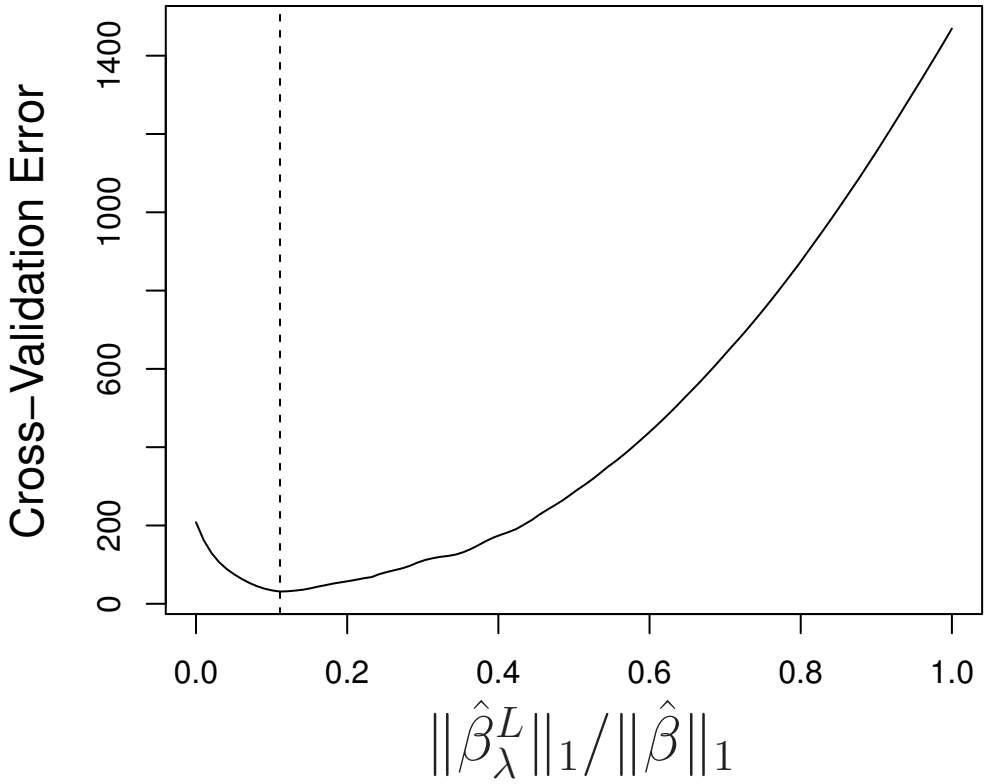
- Similar to subset selection, for ridge regression and lasso, need a method to determine which of the models under consideration is best.
  - Method to select a value for the tuning parameter $\lambda$ or equivalently, the value of the constraint $s$.
- Cross-validation provides a simple way to tackle this problem.
  - Choose a grid of $\lambda$ values, and compute the CV error rate for each value of $\lambda$.
  - Select the value of $\lambda$ for which the CV error is the smallest.
  - Refit the model using all available observations and the selected value of $\lambda$.

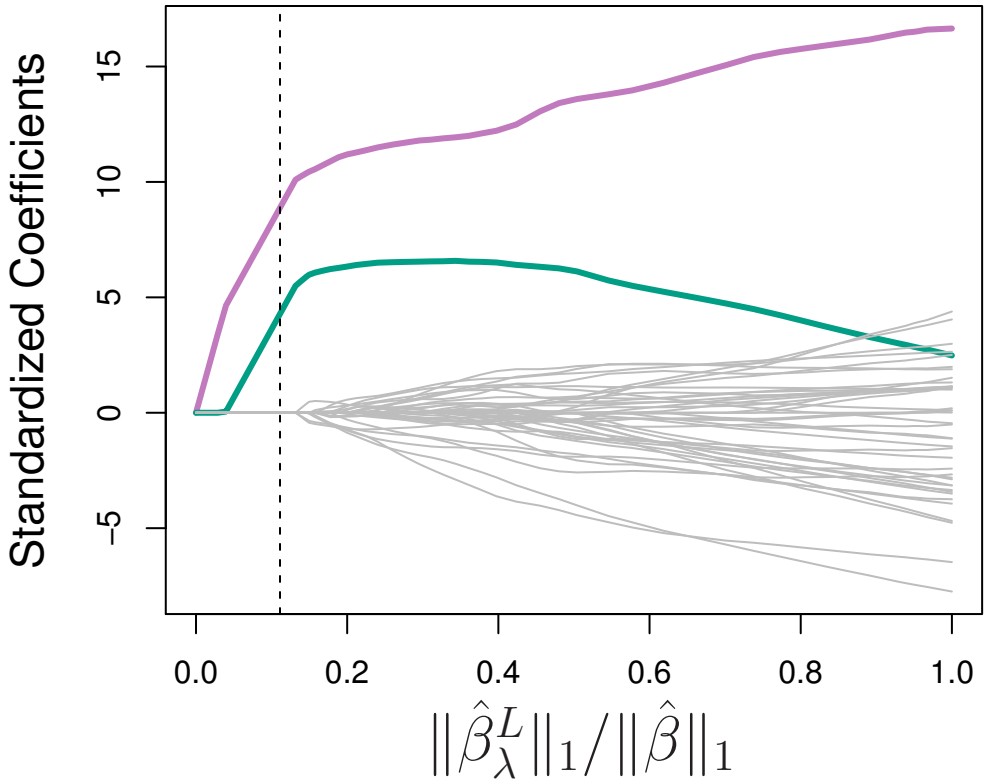
# Credit Data Example




Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various values of* $\lambda$.
Right: *The coefficient estimates as a function of* $\lambda$. *The vertical dashed lines indicates the value of* $\lambda$ *selected by cross-validation.*

# Simulation 2: Sparse Case, 2 Relevant



# Dimension Reduction Methods

- The methods that we have discussed so far have involved fitting linear regression models, via least squares or a shrunken approach, using the original predictors, $X_1, X_2, \ldots, X_p$.

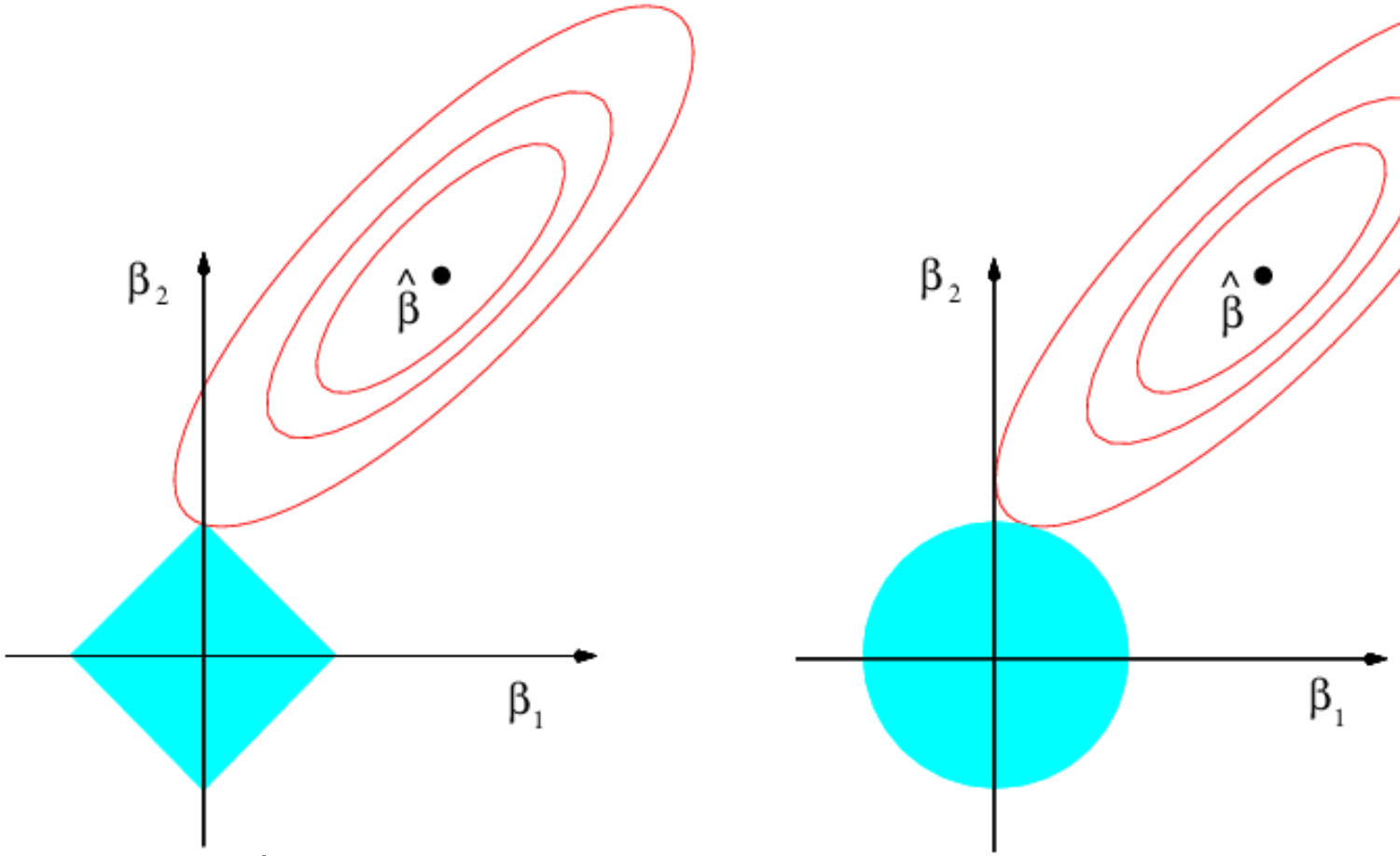
- There exist a class of approaches that transform the predictors and then fit a least squares model using the transformed variables.

- We will refer to these techniques as dimension reduction-based regression methods. (to be discussed)
  - Principal Component Regression
  - Partial Least Squares

# Summary

- Model selection methods are an essential tool for data analysis, especially for big datasets involving many predictors.
- Research into methods that give sparsity, such as the lasso is an especially hot area.
- There are other sparsity related approaches such as the elastic net.
  - Regularization via $L_0$, $L_1$, $L_2$, $L_q$, or combinations of them
  - Elastic net: $L_1 + L_2$

# The Geometry of Ridge and Lasso




$\hat{\beta}$: the least square estimate
red ellipse: region of constant RSS (increasing)
blue region: lasso/ridge constraint (driven by s)

# Stein's Paradox

- Gaussian mean problem:

$$Y_i \sim N(\theta_i, \sigma^2) \text{ for } i = 1, \ldots, n$$

  How do we estimate $\theta_i$ from $Y_i$?

- Naïve estimator: $\hat{\theta}_i = Y_i$ for all $i$.

- Can a shrinkage estimator be better than $\hat{\theta} = Y$?

- Stein's paradox: when $n \geq 3$, we have a better $\hat{\theta}$ than $\hat{\theta} = Y$ in terms of a smaller risk $E[||\hat{\theta} - \theta||^2]$. So $\hat{\theta} = Y$ is never a good choice!

# Bias and Variance of Stein's Risk

- Let $\hat{\theta} = (1-a)Y$

$$E\left[||\hat{\theta}-\theta||^2\right] = E\left[||\hat{\theta}-E\left[\hat{\theta}\right]+E\left[\hat{\theta}\right]-\theta||^2\right]$$
$$= E\left[||\hat{\theta}-E\left[\hat{\theta}\right]||^2\right] + E\left[||E\left[\hat{\theta}\right]-\theta||^2\right]$$
$$= variance + bias^2 = (1-a)^2 n\sigma^2 + a||\theta||^2.$$

- To minimize the risk, the best $a = \frac{\sigma^2 n}{\sigma^2 n + ||\theta||^2} > 0$ (Shrinkage is better!).

- James-Stein estimator: $\hat{\theta} = \left(1 - \frac{\sigma^2(n-2)}{||Y||^2}\right) Y$.
- Stein showed that JS estimator dominates $\hat{\theta} = Y$ when $n \geq 3$.

- Practical Takeaway: when we have a lot of similar things to estimate, consider shrinkage.

# Ridge Regression

- Equivalent linear regression model:

$$Y = X\theta + \varepsilon, \qquad X = I_n, \qquad \varepsilon \sim N(0, \sigma^2 I_n).$$

- Ridge regression: $\hat{\theta} = \frac{1}{1+\lambda} Y$.

# Bayesian Interpretation of Ridge and Lasso

- Recall the linear model with independent and normal errors

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

- Denote $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ with a prior distribution

$$p(\beta) = \prod_{j=1}^{p} g(\beta_j)$$

- If $g(\cdot) \sim Gaussian(0, \lambda)$, then the posterior mode (also mean) for $\beta$, given the data, is the Ridge solution.

- If $g(\cdot) \sim Laplace(0, \lambda)$, then the posterior mode for $\beta$, given the data, is the Lasso solution.

# Geometry of Bayesian Interpretation

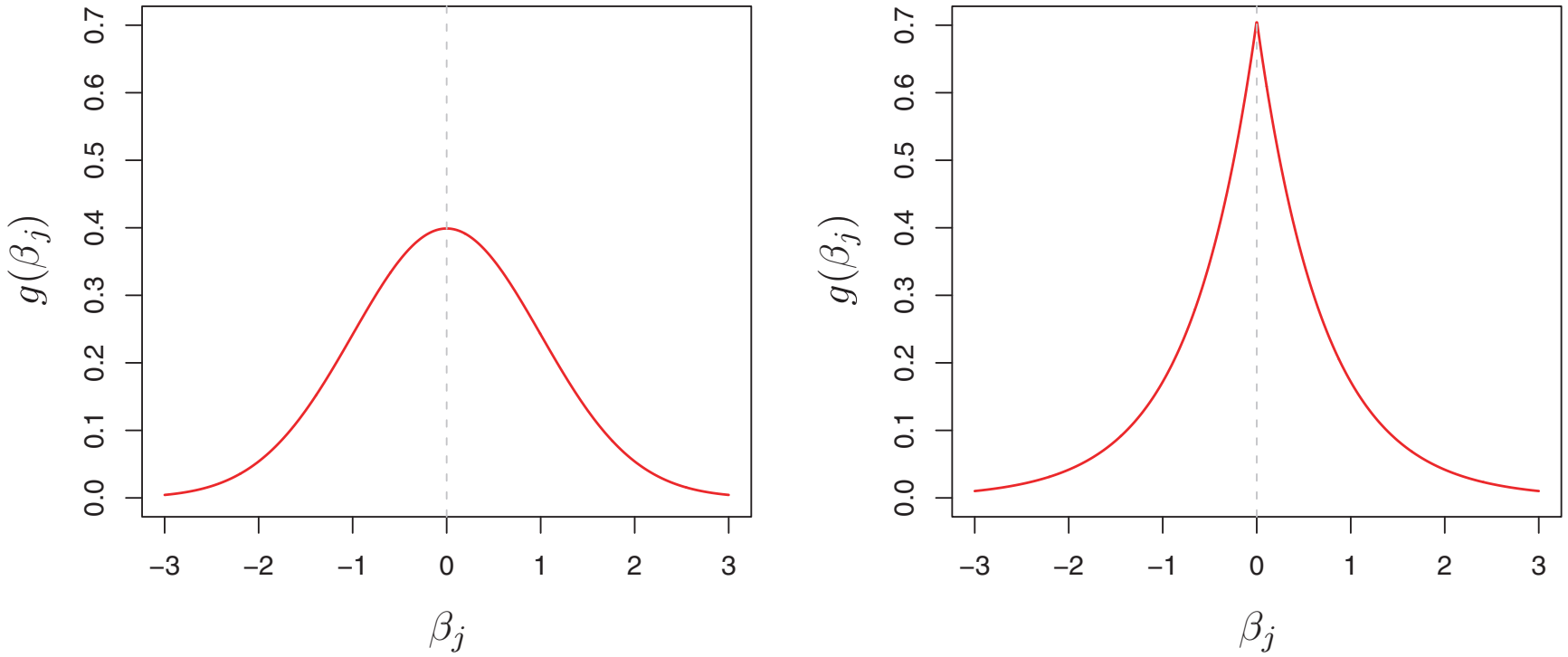



**FIGURE 6.11.** Left: *Ridge regression is the posterior mode for $\beta$ under a Gaussian prior.* Right: *The lasso is the posterior mode for $\beta$ under a double-exponential prior.*