Linear model selection and regularization

Szu-Chi Chung

Department of Applied Mathematics, National Sun Yat-sen University

Linear Model Selection and Regularization

- Recall the linear model (Can also apply to GLM)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- In the lectures that follow, we consider some approaches for extending the linear model framework. In the lectures covering Chapter 7 of the text, we generalize the linear model in order to accommodate non-linear, but still additive, relationships
- In the lectures covering Chapter 8 and 9 we consider even more general non-linear models

In praise of linear models!

- Despite its simplicity, the linear model has distinct advantages in terms of its *interpretability* and often shows good *predictive performance*
  - Hence we discuss in this lecture some ways in which the simple linear model can be improved, by replacing ordinary least squares fitting with some alternative fitting procedures
  - This often applies to the case when $n \approx p$ or $p > n$

Why consider alternatives to least squares?

1. **Prediction Accuracy**: especially when $p > n$ or $p \approx n$, to control the variance
2. **Model Interpretability**: By removing *irrelevant* or *redundant* features – that is, by setting the corresponding coefficient estimates to zero – we can obtain a model that is more easily interpreted. We will present some approaches for automatically performing feature selection
3. **Speed up** the training/inference
4. Avoid the **curse of dimensionality**

Three classes of methods

1. **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
2. **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
3. **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

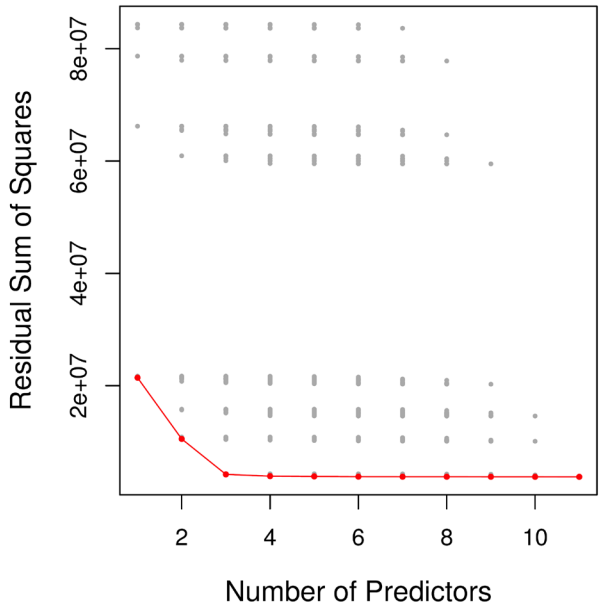
1. Subset Selection

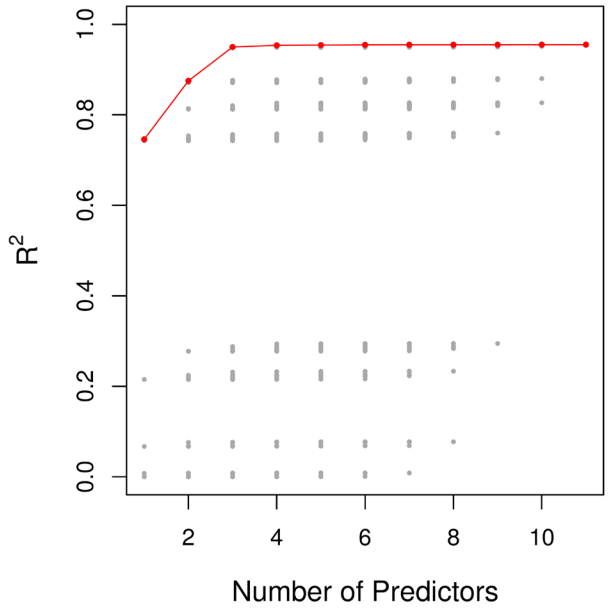
Best subset selection procedures

1. Let 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 M_k. Here best is defined as having the smallest RSS, or equivalently largest R^2
3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2

Example - Credit card data set

- For each possible model containing a subset of the ten predictors in the credit card dataset, the RSS and R^2 are displayed
  - 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
  - The red frontier tracks the best model for a given number of predictors





Extensions 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
- 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: 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
  - Thus an enormous search space can lead to overfitting and high variance of the coefficient estimates
- For both of these reasons, stepwise methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection

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

In Detail

Forward Stepwise Selection

1. Let 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 M_k with one additional predictor
   b) Choose the best among these $p-k$ models, and call it M_{k+1}. Here best is defined as having the smallest RSS, or equivalently largest R^2
3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2

Though forward stepwise selection considers $p(p+1)/2+1$ models, it performs a guided search over model space, and so the effective model space considered contains substantially more than $p(p+1)/2+1$ models

More on Forward Stepwise Selection

- ✓ The computational advantage over best subset selection is clear
  - For high dimensional data with $p > n$, the forward selection can still be applied by considering only $M_1, \ldots, M_n$, since each submodel fit with least square will not have unique solution for $p > n$
- x It is not guaranteed to find the best possible model (lowest training error) out of all 2^p models containing subsets of the p predictors
  - Suppose that in a given data set with $p = 3$ predictors, the best possible one-variable model contains X_1, and 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

Credit data example

- 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
  - In this example, there is actually not much difference between the three and four-variable models in terms of RSS, so either of the four-variable models will likely be adequate

# 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

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

Backward Stepwise Selection

1. Let 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 M_k , for a total of $k-1$ predictors
   b) Choose the best among these k models, and call it M_{k-1}. Here best is defined as having the smallest RSS, or equivalently largest R^2
3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2

More on Backward Stepwise Selection

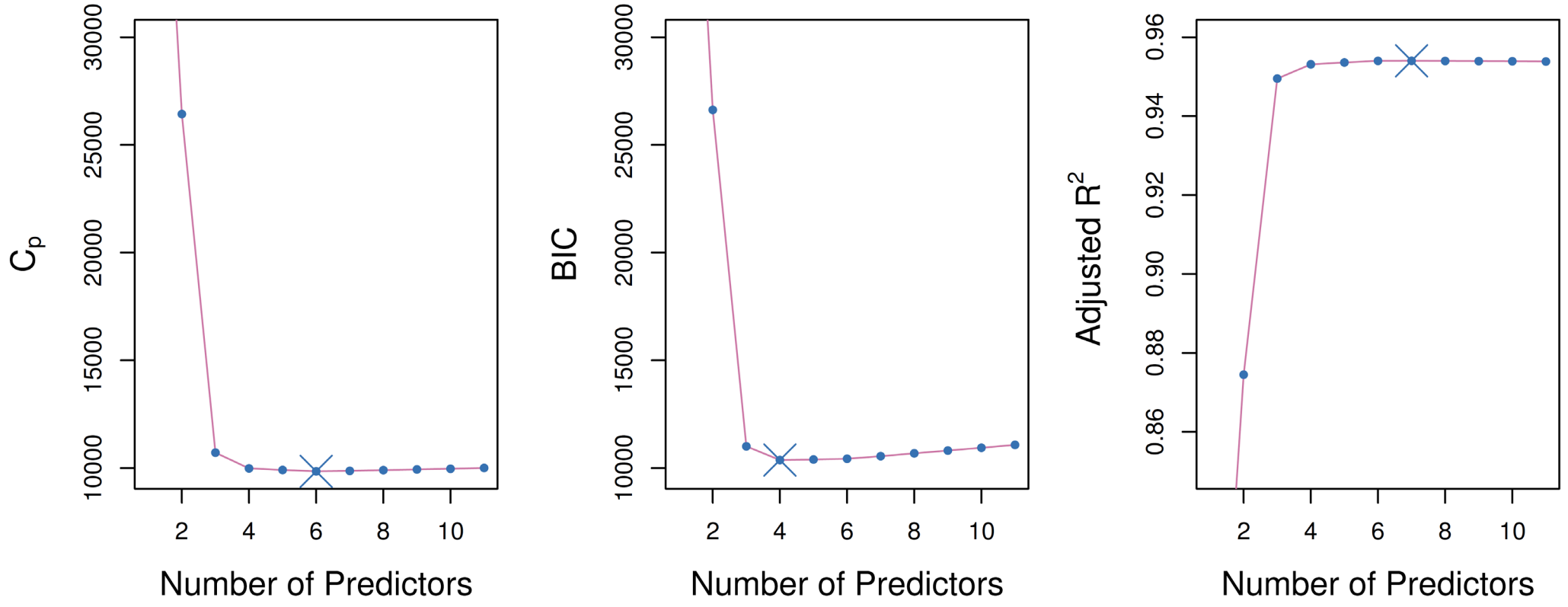
- ✓ Like forward stepwise selection, the backward selection approach searches through only $1 + p(p + 1)/2$ models, and so can be applied in settings where p is too large to apply the best subset selection
- x Backward selection requires that $n \geq p$ (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large
- x Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best (lowest training error) model containing a subset of the p predictors

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 numbers of predictors
- Therefore
  - We can directly estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in previous lectures
  - We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting

Credit data example

- The figure displays C_p (AIC), BIC, and adjusted R^2 for the best model of each size produced by best subset selection on the Credit card data set



Details on C_p and AIC

- Mallow's C_p for estimated test MSE (for least square model):

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)\ , \hat{\sigma}^2 = \frac{\widehat{RSS}}{n-p-1}$$

  - where d is the total # of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement based on model *containing all predictors*

- The AIC (Akaike Information Criterion) is defined for a large class of models fit by maximum likelihood:

$$AIC = -2\log L + 2 \cdot d = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) + Const$$

  - Where L is the maximized value of the likelihood function for the estimated model
  - In the case of the linear model with Gaussian errors, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent

Details on BIC

- BIC (Bayesian information criterion) is motivated in quite a different way. It arises in the Bayesian approach to model selection

$$BIC = -2\log L + \log(n)\, d = \frac{1}{n}(RSS + \log(n)\, d\hat{\sigma}^2) + Const$$

  - Like C_p, the BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value
  - Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)\, d\hat{\sigma}^2$ 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

Adjusted R^2

- For a least squares model with d variables, the adjusted R^2 statistic is calculated as

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \left(R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}\right)$$

where TSS$= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the total sum of squares

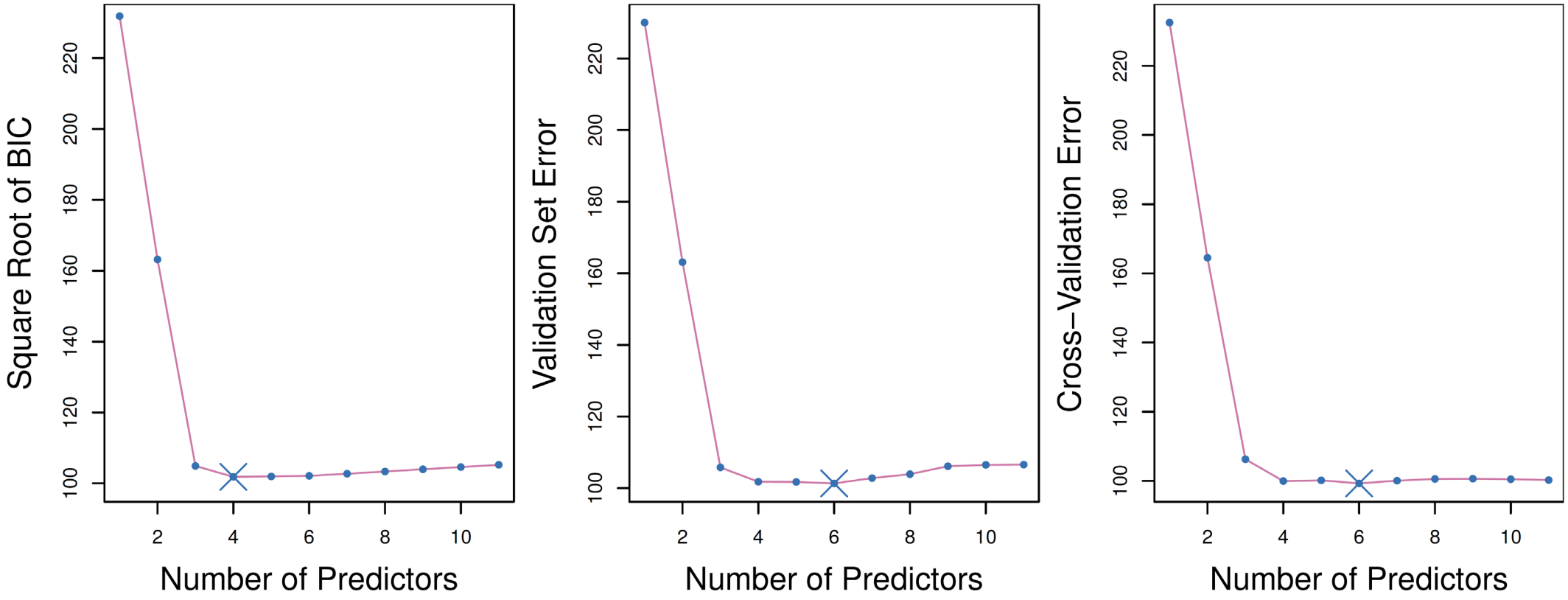
  - Maximizing the adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-d-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{RSS}{n-d-1}$ may increase or decrease, due to the presence of d in the denominator
  - Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model
  - Unlike C_p, AIC, and BIC, for which a small value indicates a model with a low test error, a large value of adjusted R^2 indicates a model with a small test error

Validation and cross-validation

- Each of the procedures returns a sequence of models M_k indexed by model size $k = 0,1,2 \ldots$. Our job here is to select $\hat{k}$. Once selected, we return model $M_{\hat{k}}$
  - We compute the validation set error or the cross-validation error for each model 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: 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 σ^2
  - x It needs computational power

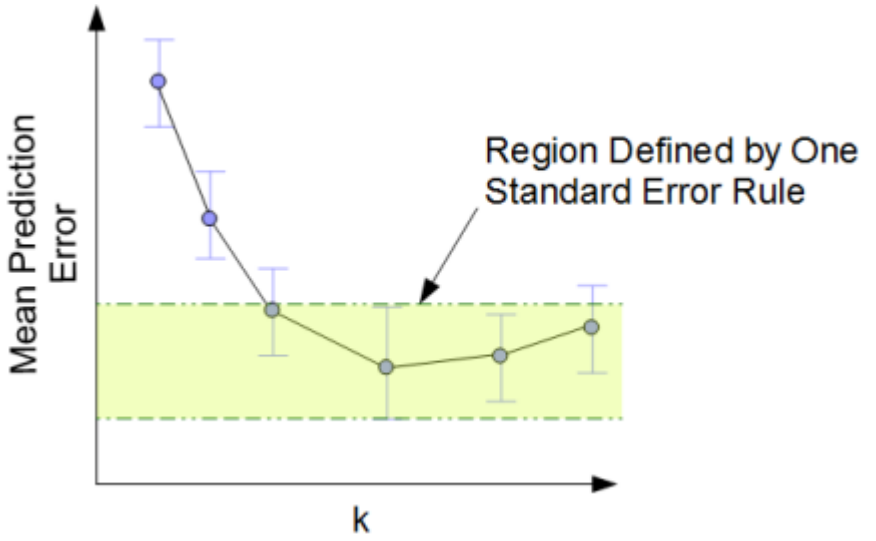
Credit data example

- 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



One-standard-error rule

- All three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors
  - In this setting, we can select a model using the one-standard-error rule. We first calculate the standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve



https://www.cs.cmu.edu/~psarkar/sds383c_16/lecture9_scribe.pdf

2. Shrinkage Methods

- Here we will discuss about 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

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

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

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

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

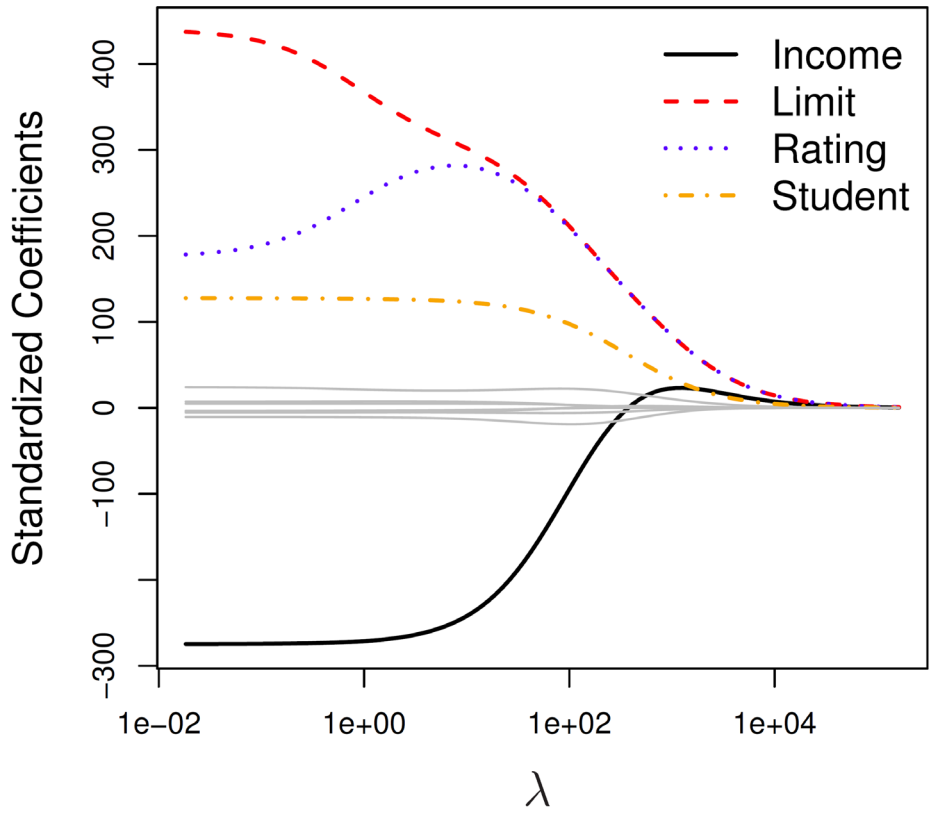
where $\lambda \geq 0$ is a tuning parameter, to be determined separately

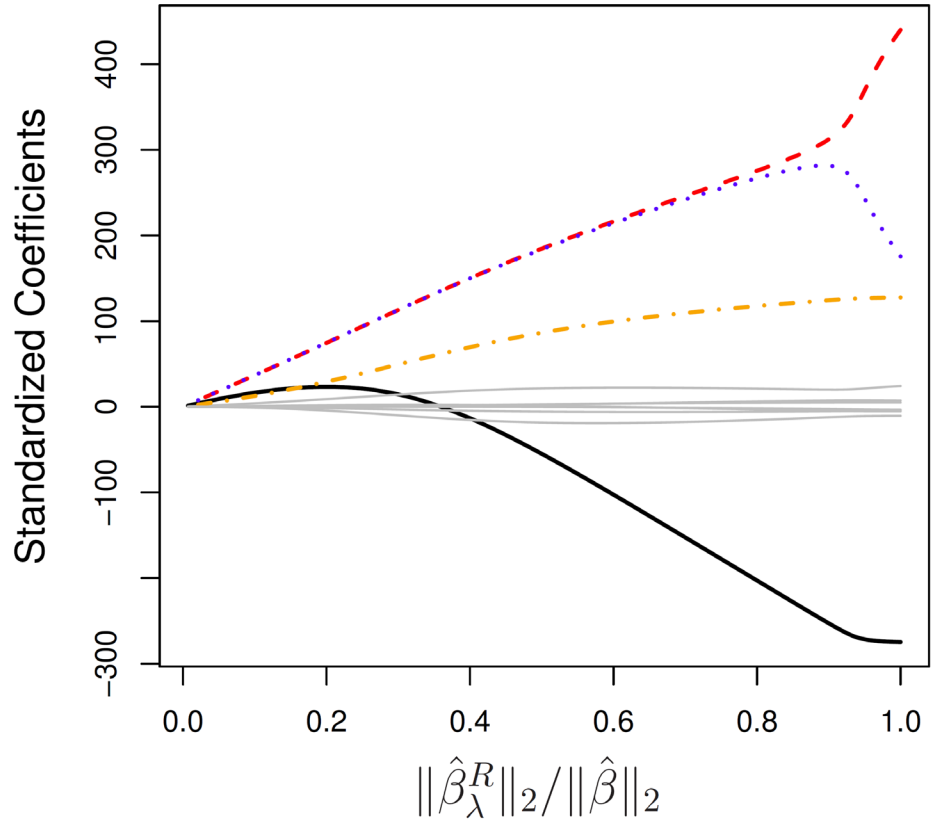
Ridge regression: continued

- 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 β_j towards zero
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates
  - Selecting a good λ value for is critical; cross-validation is used for this

Credit data example

- For all 10 features





Details of Previous Figure

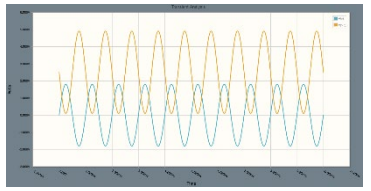
- In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but instead of displaying on the x-axis, we now display $\left\|\hat{\beta}_\lambda^R\right\|_2 / \left\|\hat{\beta}\right\|_2$ (0~1, *shrinkage factor*),where $\hat{\beta}$ denotes the vector of least squares coefficient estimates
  - The notation $\|\beta\|_2$ denotes the l_2 norm (pronounced "ell 2") of a vector, and is defined as $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

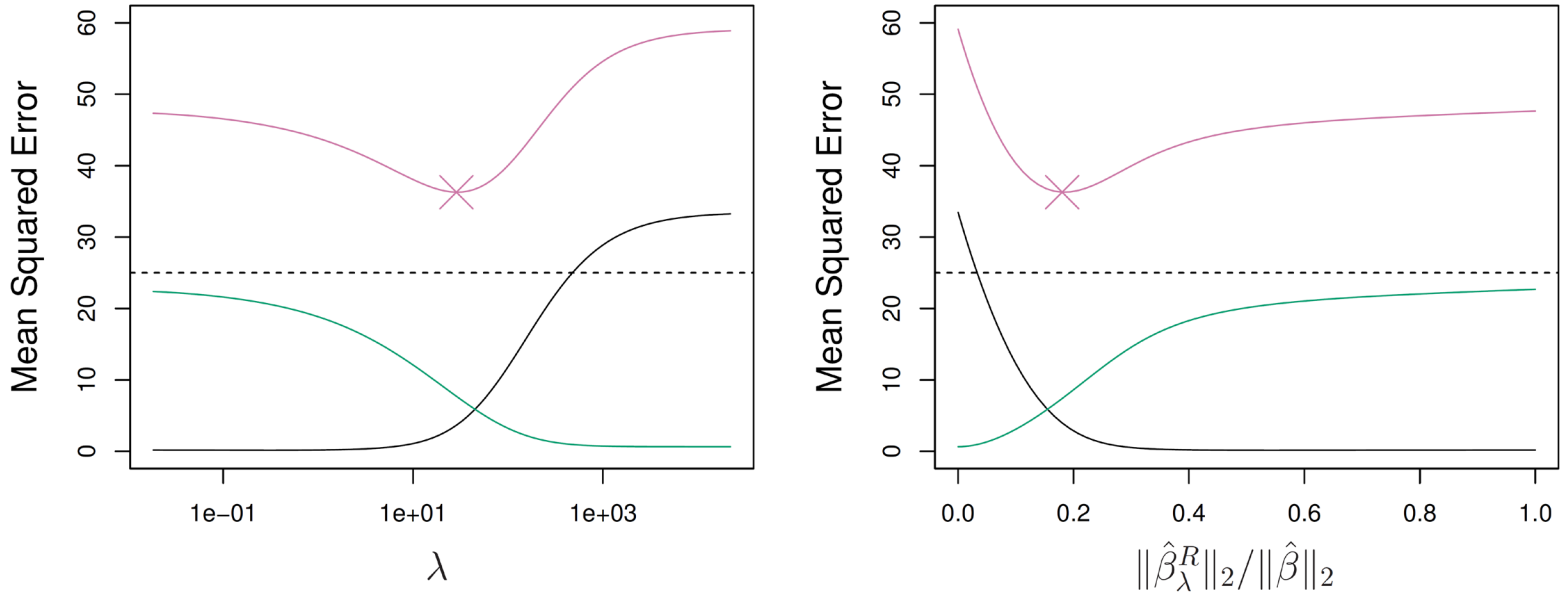
Ridge regression: scaling of predictors

- The standard least squares coefficient estimates are scale equivalent: 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 jth 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 (Same for the PLS in the following slides)

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

Why Does Ridge Regression Improve 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 λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. 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 (Least Absolute Shrinkage and Selection Operator)

- ✗ 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

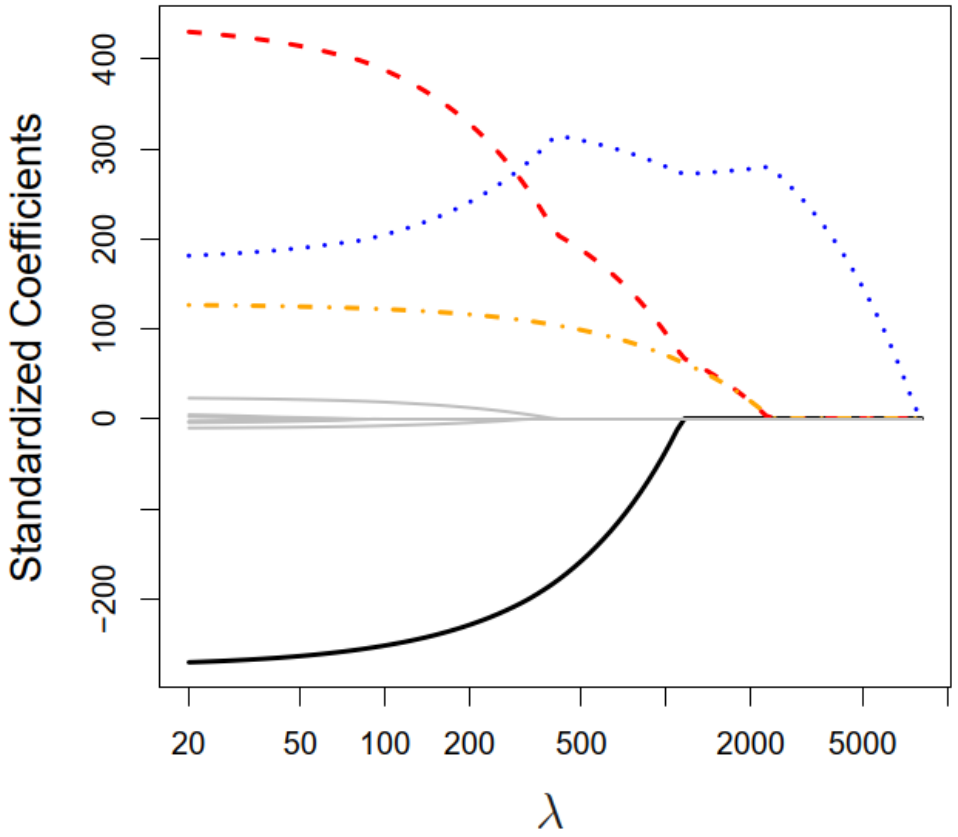
$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

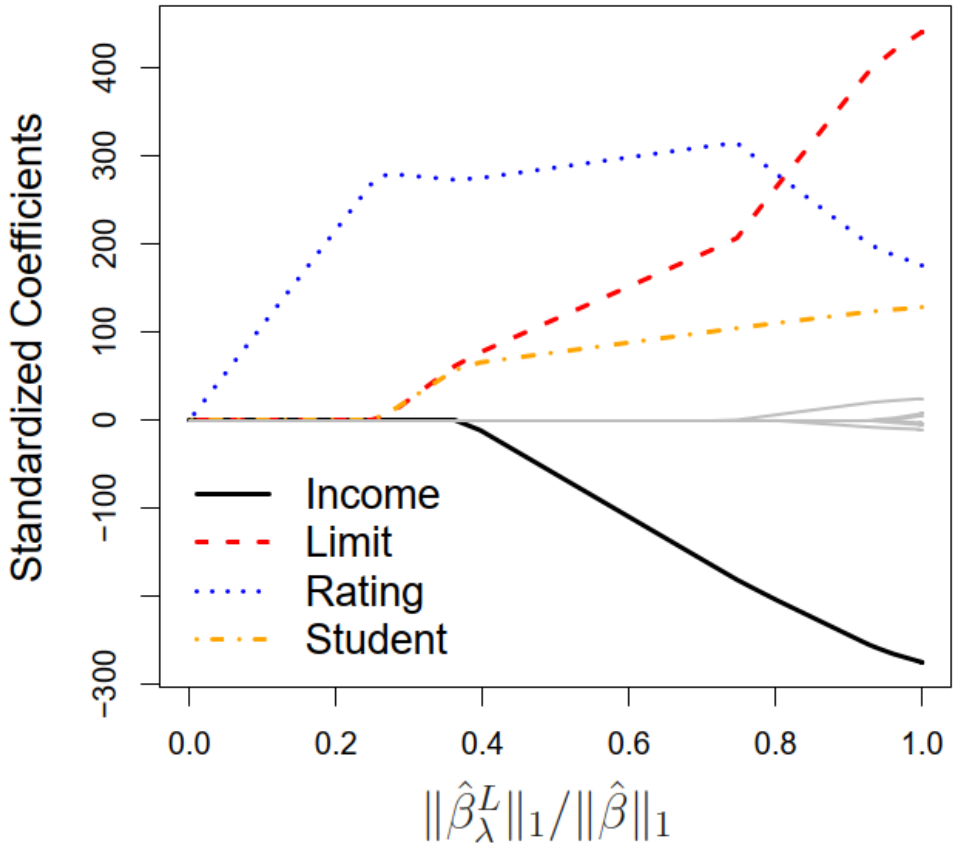
In statistical parlance, the lasso uses an l_1 norm (pronounced "ell 1") penalty instead of an l_2 penalty. The l_1 norm of a coefficient vector β is given by $\|\beta\|_1 = \sum|\beta_j|$

The Lasso: continued

- As ridge regression, the lasso shrinks the coefficient estimates towards zero
  - However, in the case of the lasso, the l_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ 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 λ for the lasso is critical; cross-validation is again the method of choice

Example: Credit dataset





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 (linked by the Lagrange multiplier)

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

and

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

Respectively

- The best subset selection can be viewed as (ESL, ch3)

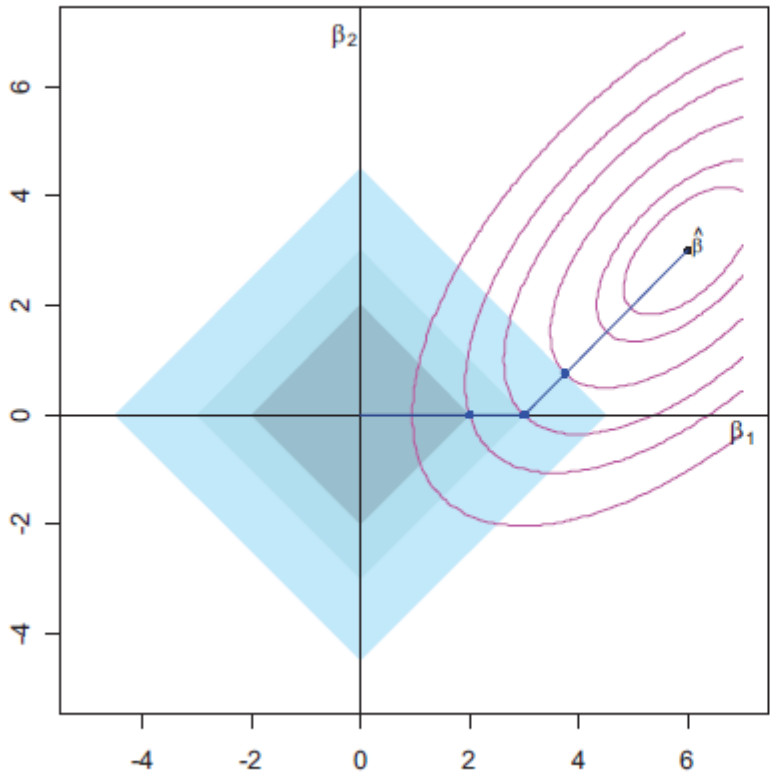
$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \text{ subject to } \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s$$

Solving above formula is, however, computationally infeasible

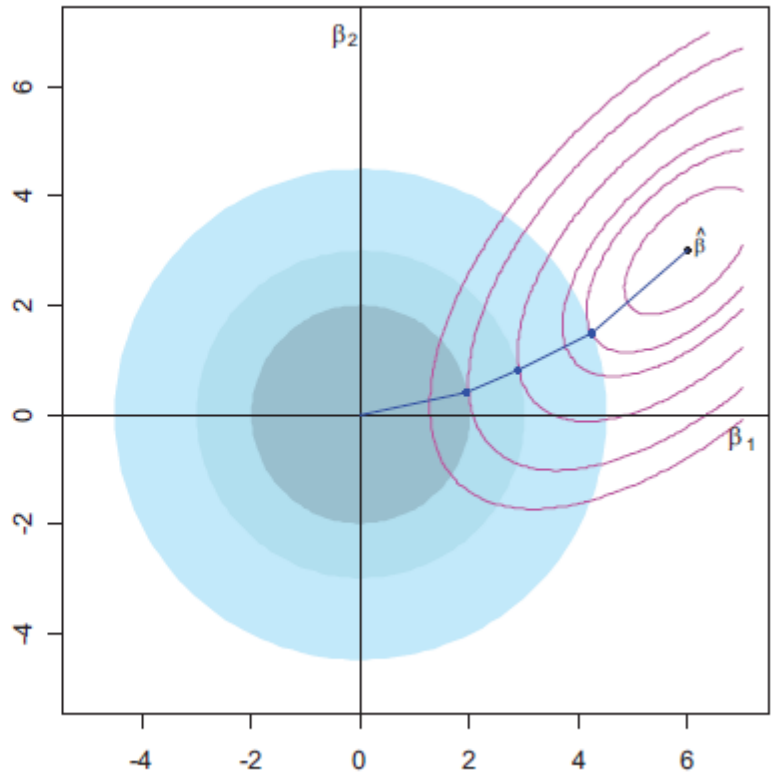
The Lasso Picture

contour of RSS

Lasso



Ridge



(budget) constraint by s

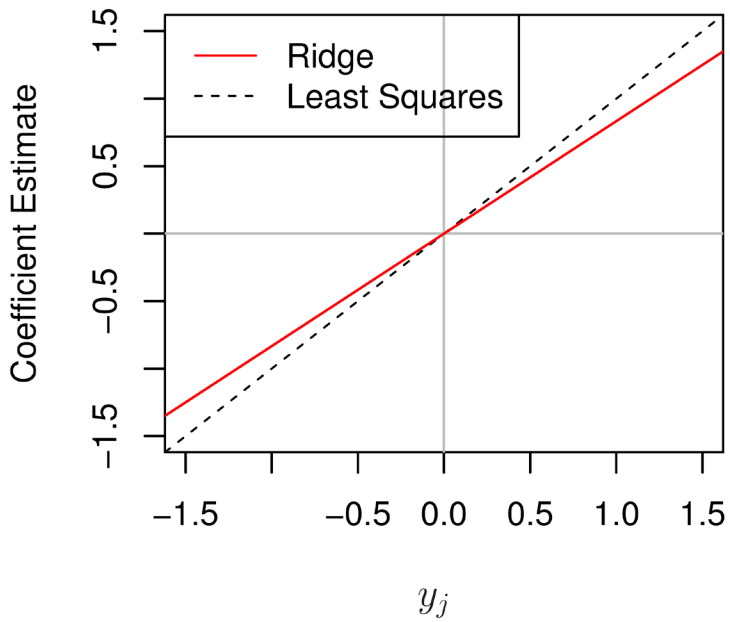
https://stats.stackexchange.com/questions/350046/the-graphical-intuiton-of-the-lasso-in-case-p-2?noredirect=1&lq=1

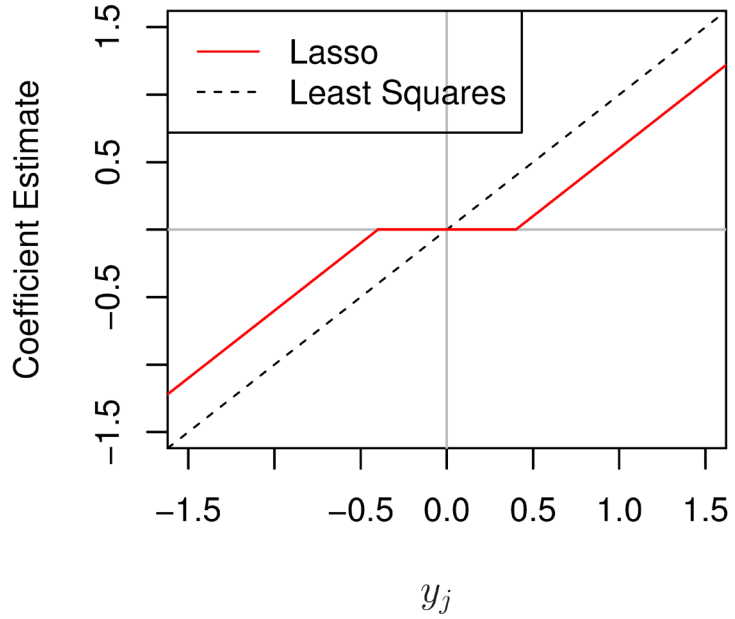
More about intuition (exercise 6)

- Consider X is a square diagonal matrix with its diagonal elements equal to 1 and we omit the intercept for simplicity
  - The least squares problem in this case is to minimized $\sum_{j=1}^{p}(y_j - \beta_j)^2 \rightarrow \hat{\beta}_j = y_j$
  - Ridge regression: $\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \rightarrow \hat{\beta}_j^R = y_j/(1+\lambda)$
  - Lasso: $\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \rightarrow \hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & if\ y_j > \lambda/2 \\ y_j + \lambda/2 & if\ y_j < -\lambda/2 \\ 0 & if\ |y_j| \leq \lambda/2 \end{cases}$
    - Close-form solution available when features are uncorrelated for lasso

More about intuition

- 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. In the case of a more general data matrix X the main ideas still hold approximately

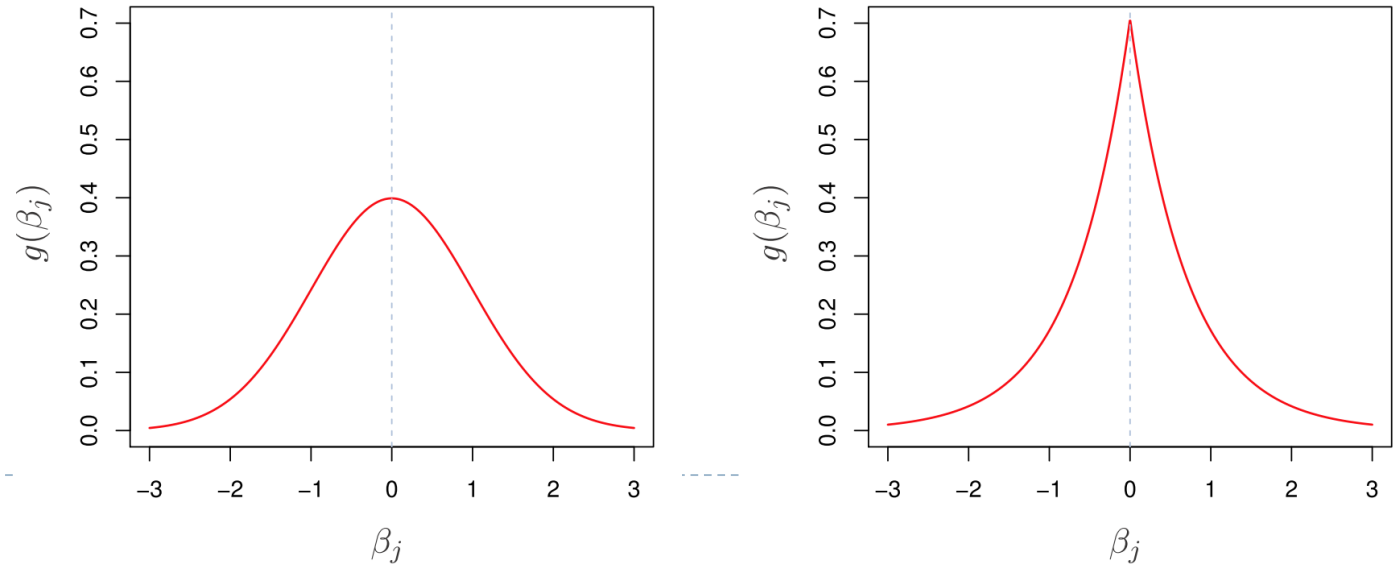




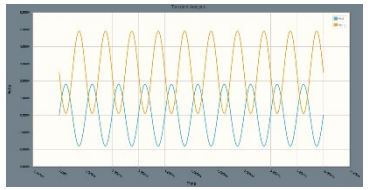
Bayesian interpretation (exercise 7)

$$p(\beta|X,Y) \propto f(Y|X,\beta)p(\beta)$$

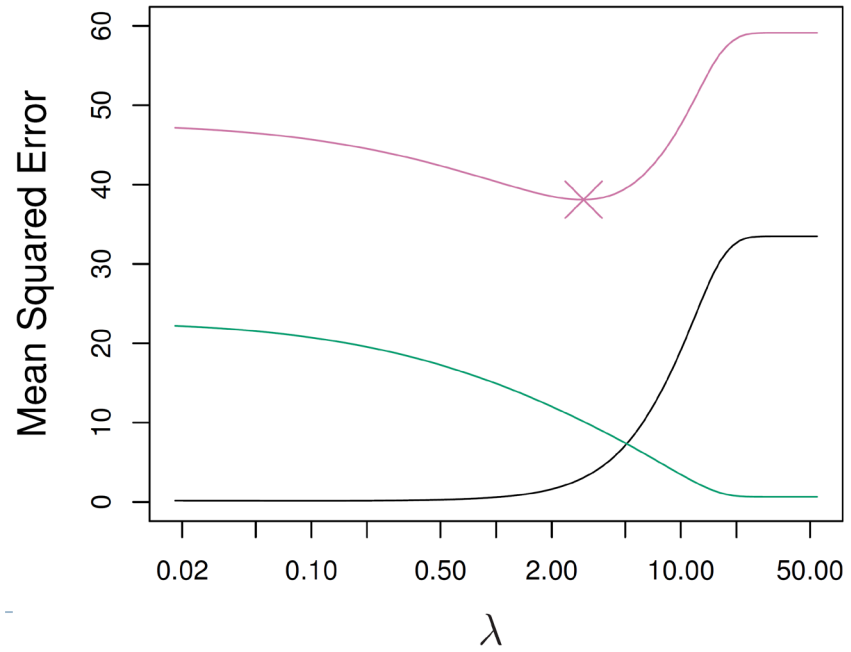
- We assume that $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$, for some density function g
- If g is a Gaussian distribution with mean zero and standard deviation a function of λ, then it follows that the posterior mode for β—that posterior is, the most likely value for β, given the data—is given by the ridge mode regression solution!
- If g is a double-exponential (Laplace) distribution with mean zero and scale parameter a function of λ, then it follows that the posterior mode for β is the lasso solution!

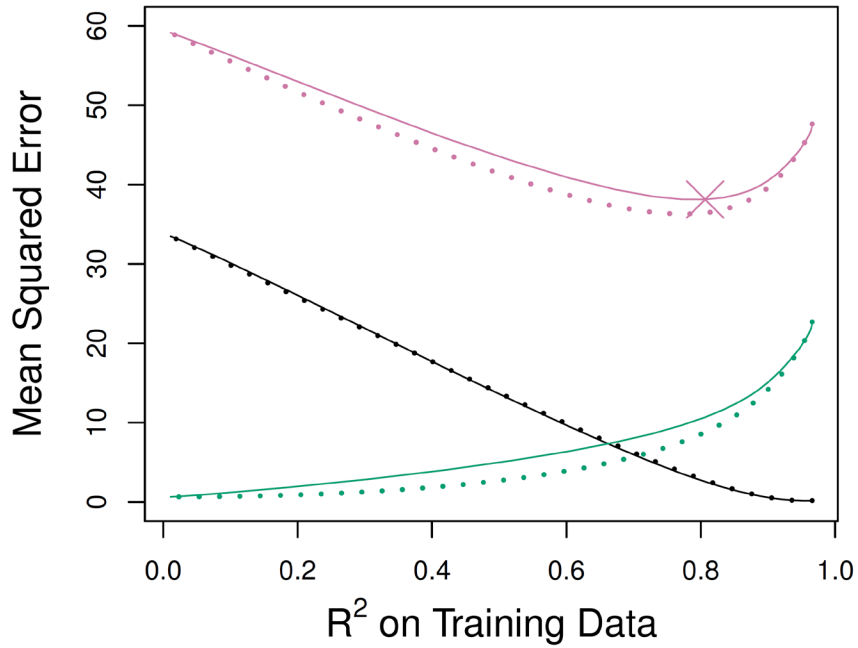


Comparing the Lasso and Ridge Regression



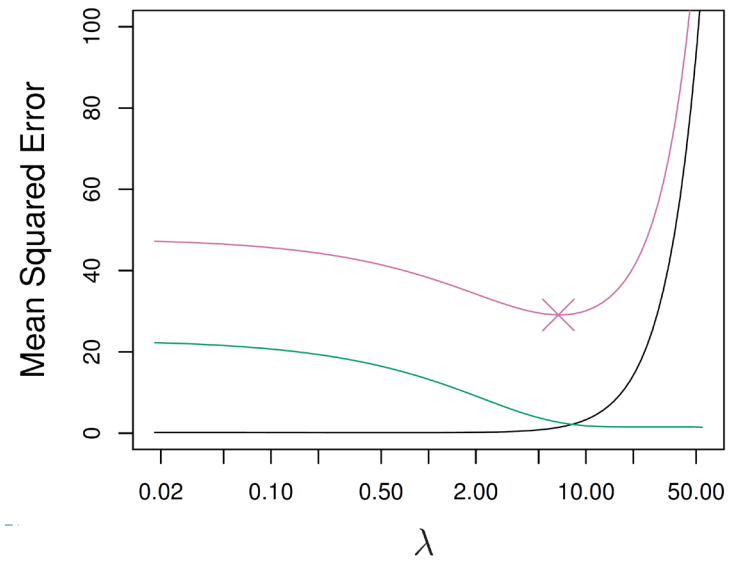
- Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on simulated data set of Slide 31
- Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). Both are plotted against their R^2 on the training data, as a common form of indexing

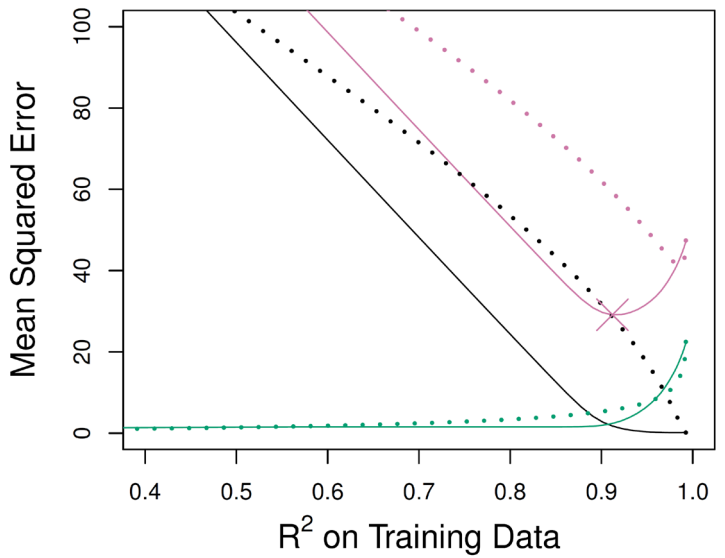




Comparing the Lasso and Ridge Regression: continued

- Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Slide 31, except that now only two predictors are related to the response
- 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





Short Conclusions

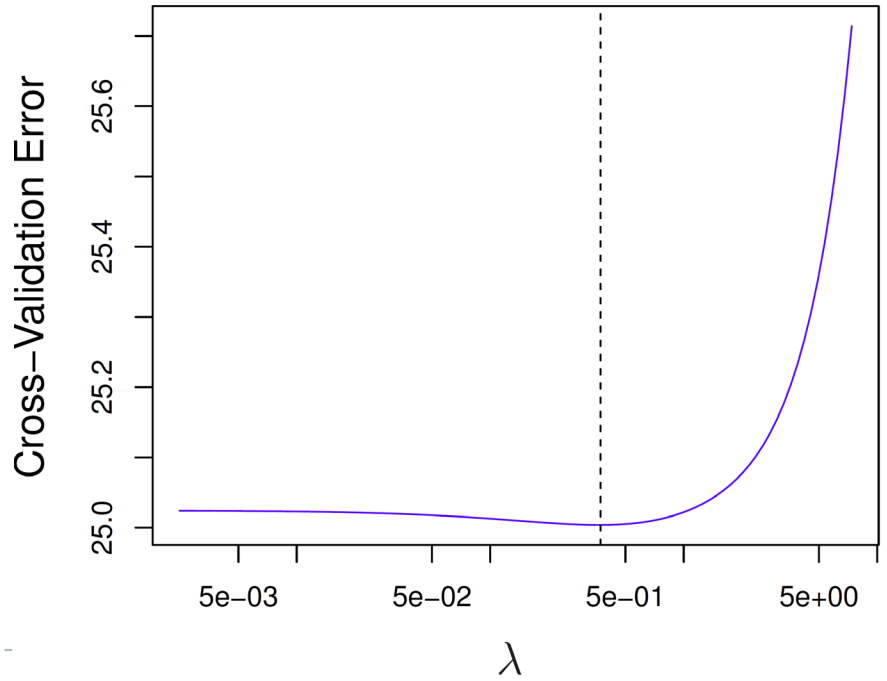
- These two examples illustrate that neither ridge regression nor the lasso will universally dominate the other
- In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors
  - However, the number of predictors that is related 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

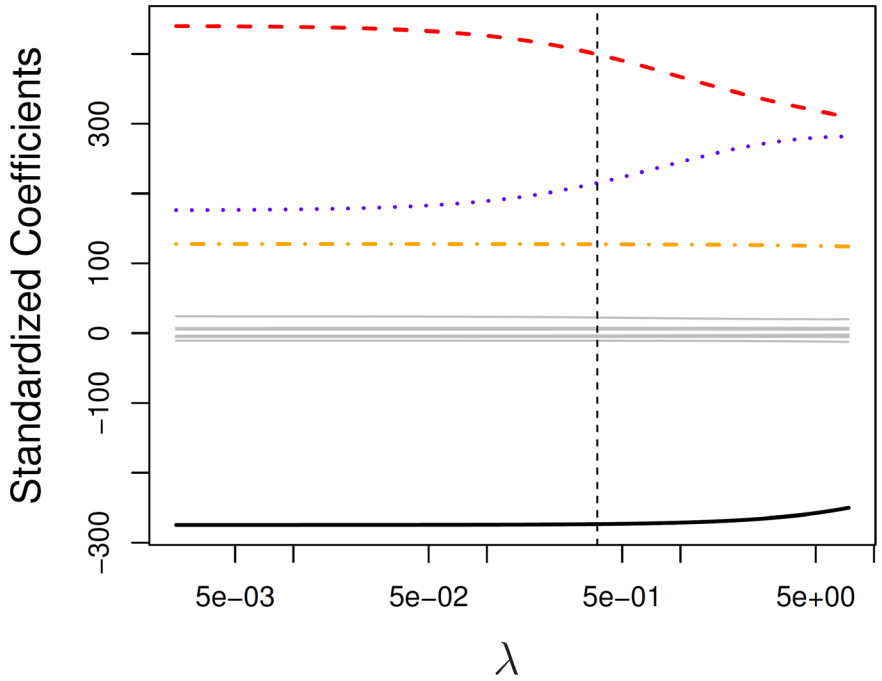
Selecting the Tuning Parameter for Ridge Regression and Lasso

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration is best
  - That is, we require a method selecting a value for the tuning parameter λ or equivalently, the value of the constraint s
- Cross-validation provides a simple way to tackle this problem. We choose a *grid of λ values*, and compute the cross-validation error rate for each value of λ
  - We then select the tuning parameter value for which the cross-validation error is smallest
  - Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter

Credit data example

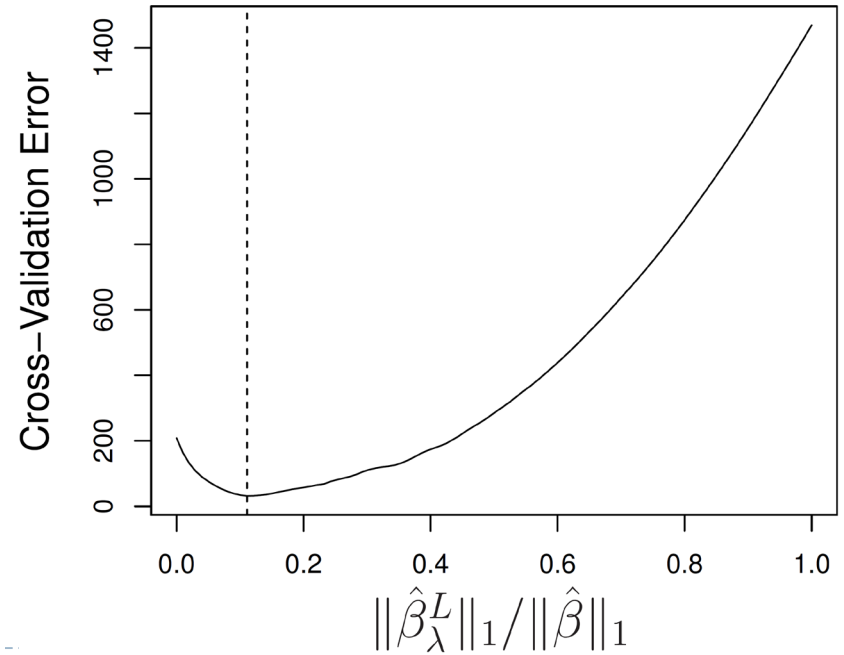
- Left: LOOCV errors that result from applying ridge regression to the Credit data set with various values of λ
- Right: The coefficient estimates as a function of λ. The vertical dashed lines indicates the value of λ selected by cross-validation

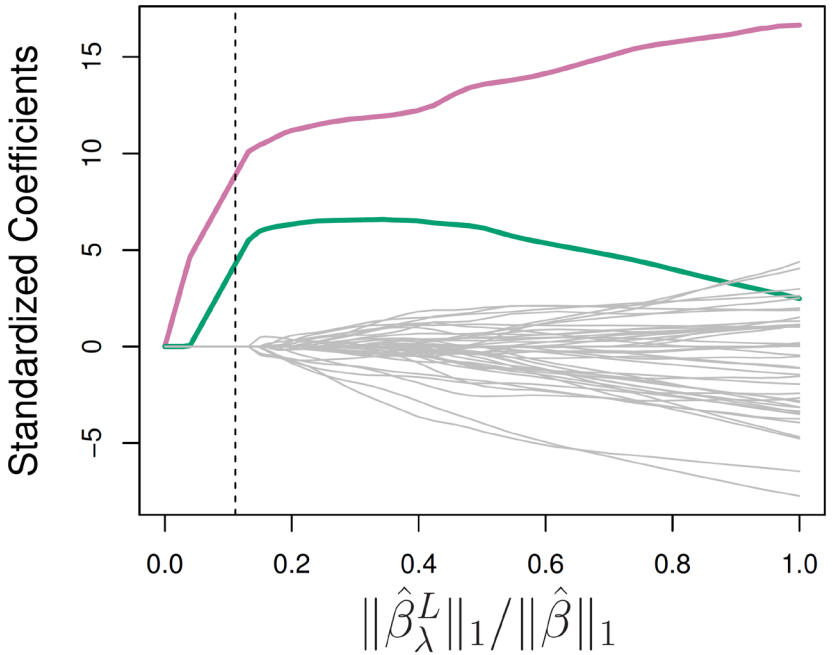




Simulated data example

- Left: Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Slide 41
- Right: The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest





3. Dimension Reduction Methods

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

Dimension Reduction Methods: details

- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ linear combinations of our original p predictors. That is,

Basis

$$Z_m = \sum_{j=1}^{p} \Phi_{jm} X_j \quad \text{for some constants } \Phi_{1m}, \ldots, \Phi_{pm}$$

  - We can then fit the linear regression model using ordinary least squares

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, i = 1, \ldots, n$$

- Note that in the model, the regression coefficients are given by $\theta_0, \theta_1, \ldots, \theta_M$. If the constants $\Phi_{1m}, \ldots, \Phi_{pm}$ are chosen wisely, then such dimension reduction approaches can often outperform OLS regression

Dimension Reduction Methods: details

- Notice that from definition

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \Phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \Phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

  Where $\beta_j = \sum_{m=1}^{M} \theta_m \Phi_{jm}$

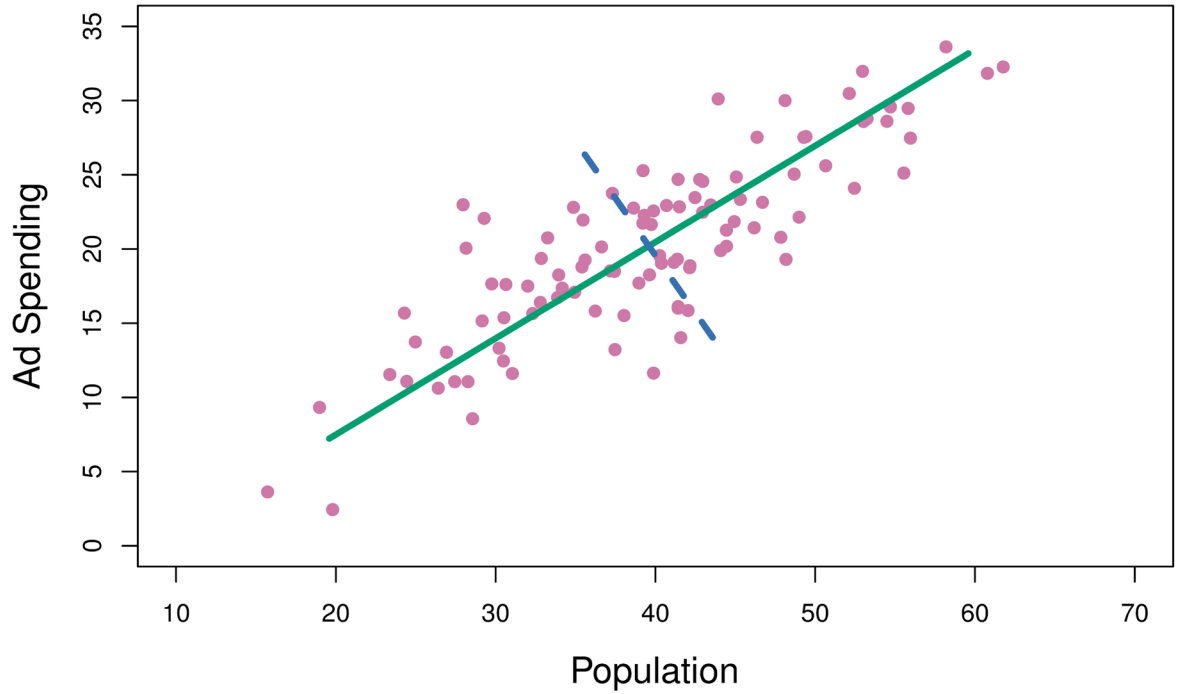
  - Hence the model can be thought of as a special case of the original linear regression model
  - Dimension reduction serves to constrain the estimated β_j coefficients, since now they must take the above form. Can win in the bias-variance tradeoff!

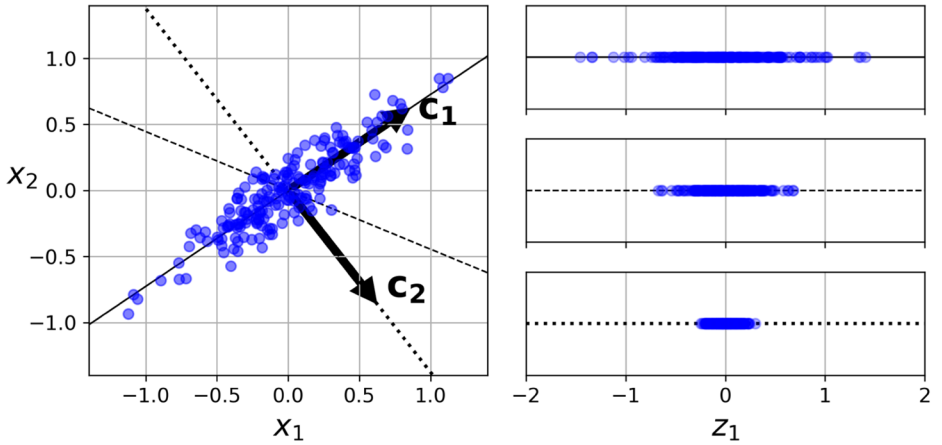
Principal Components Regression

- Here we apply principal components analysis (PCA) (discussed in Chapter 12 of the text) to define the linear combinations of the predictors
  - The first principal component is that (normalized) direction with the largest variance
  - The second principal component score has largest variance, subject to being uncorrelated with the first. And so on
  - Hence with many correlated original variables, we replace them with a small set of uncorrelated principal components scores that capture their joint variation
- The principal components regression (PCR) approach involves constructing the first M principal components scores, $Z_1, Z_2, \ldots, Z_M$, and then using these components scores as the predictors in a linear regression model that is fit using least squares

Pictures of PCA

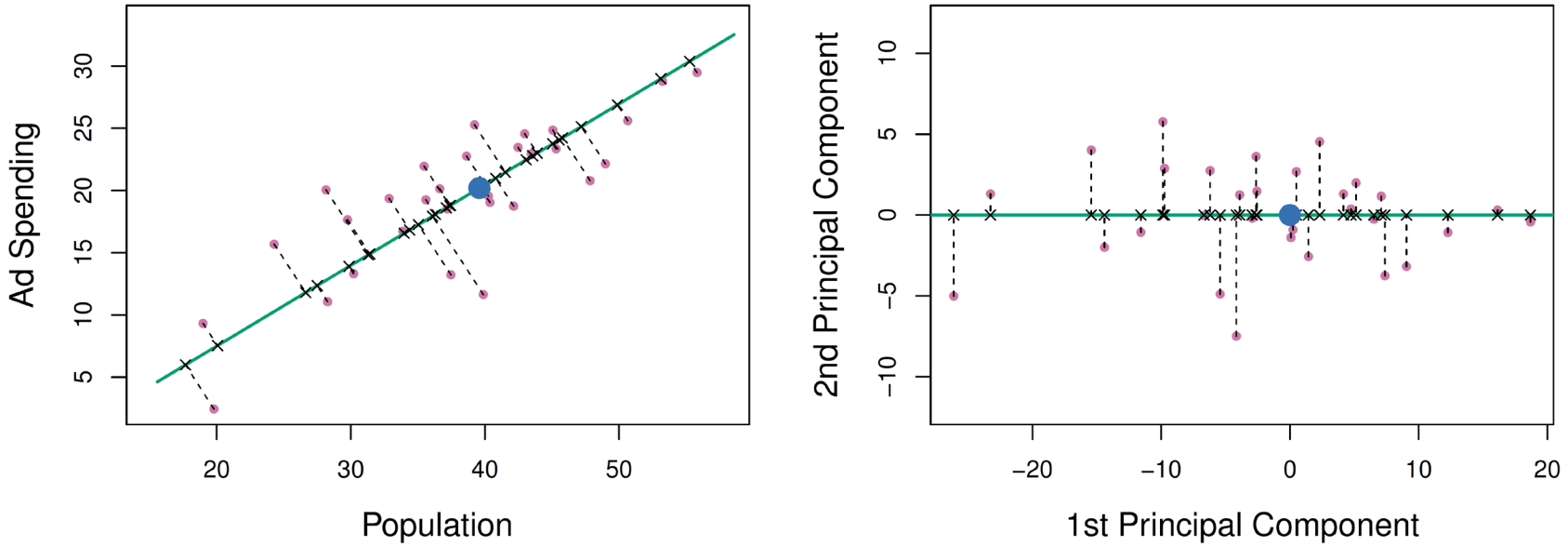
- The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component





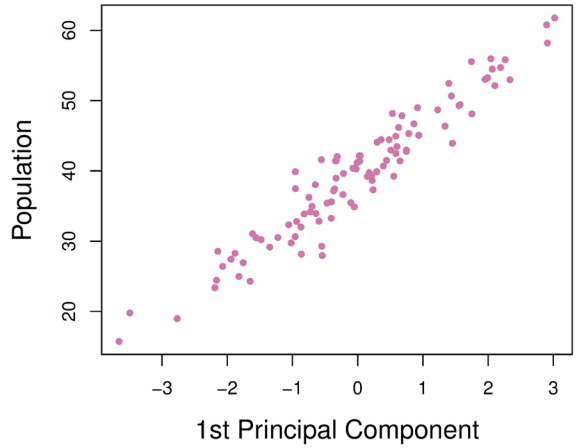
Pictures of PCA: continued

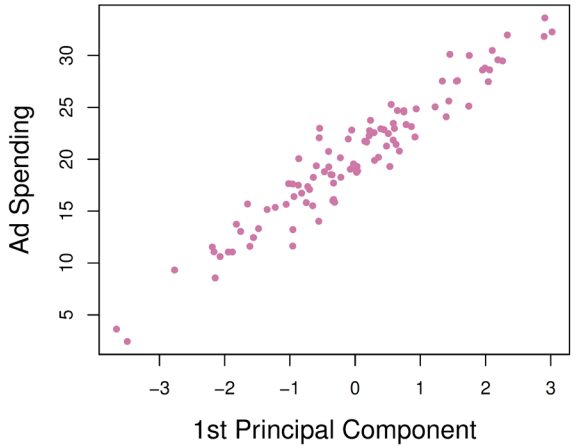
- Left: The first principal component, chosen to minimize the sum of the squared perpendicular distances to each point, is shown in green. These distances are represented using the black dashed line segments
- Right: The left-hand panel has been rotated so that the first principal component lies on the x-axis (We need to standardized the data first)



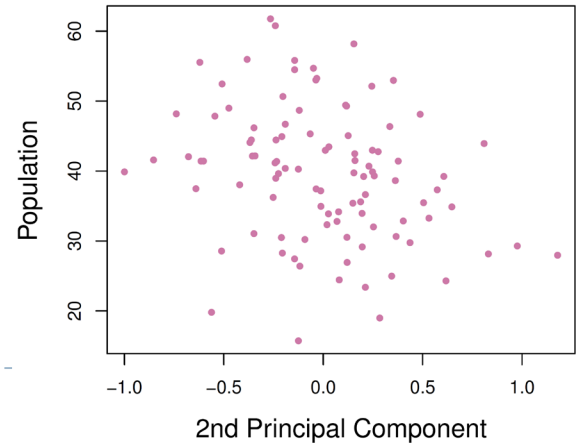
Pictures of PCA: continued

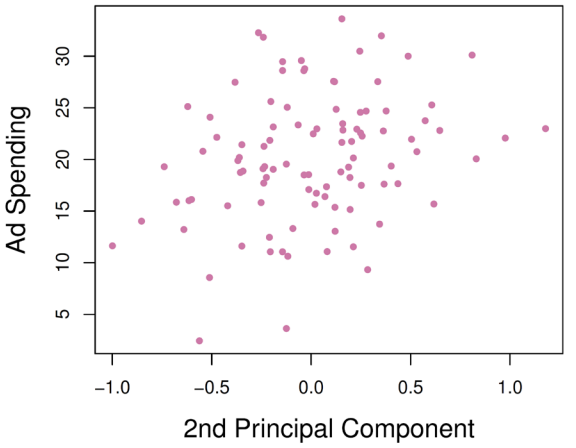
- Plots of the first principal component scores z_{i1} versus pop and ad. The relationships are strong



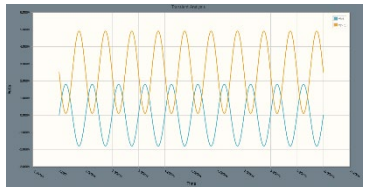


- Plots of the second principal component scores z_{i2} versus pop and ad. The relationships are weak

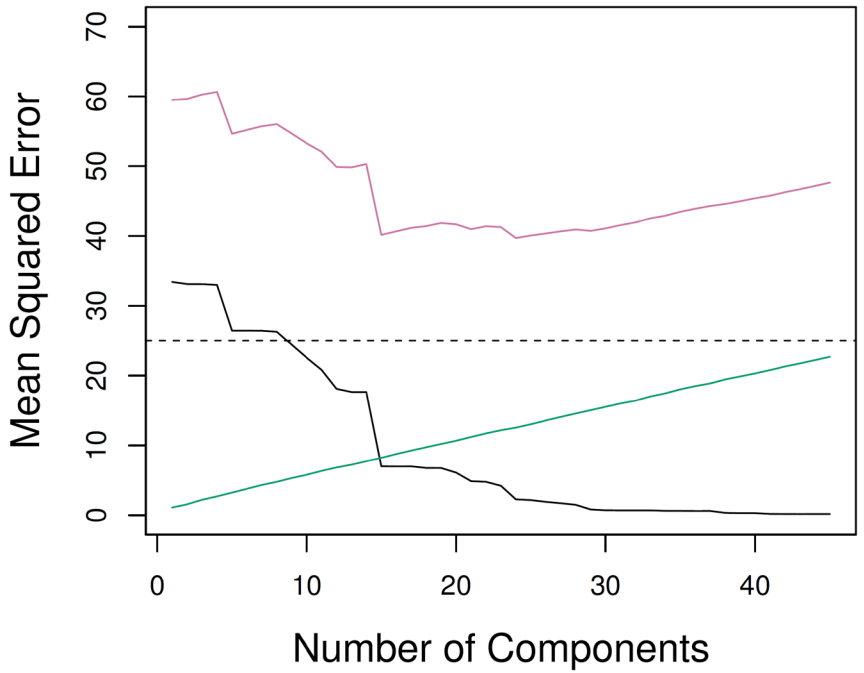




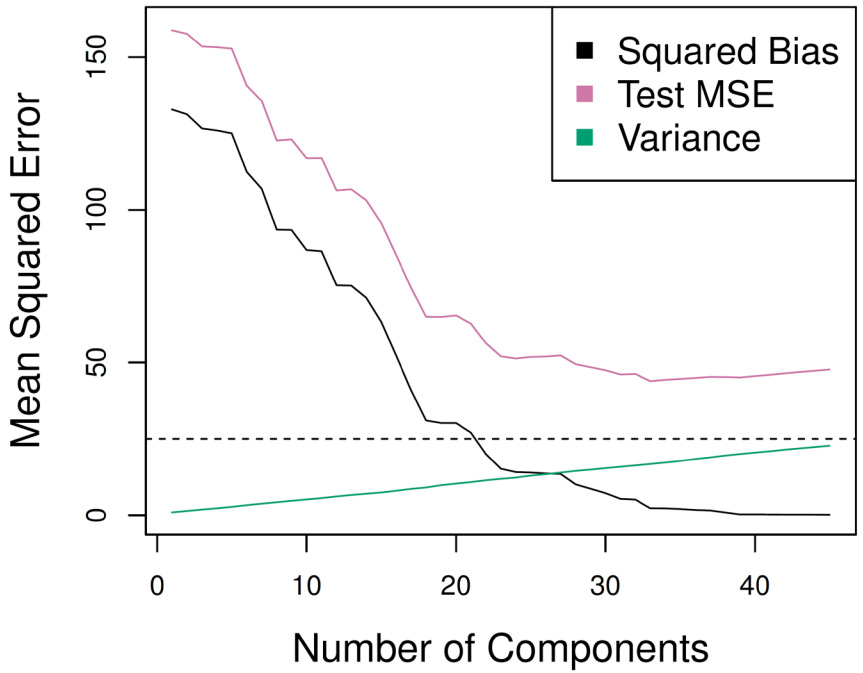
Application to Principal Components Regression



- PCR was applied to two simulated data sets. The black, green, and purple lines correspond to squared bias, variance, and test mean squared error, respectively. Left: Simulated data from slide 31. Right: Simulated data from slide 41



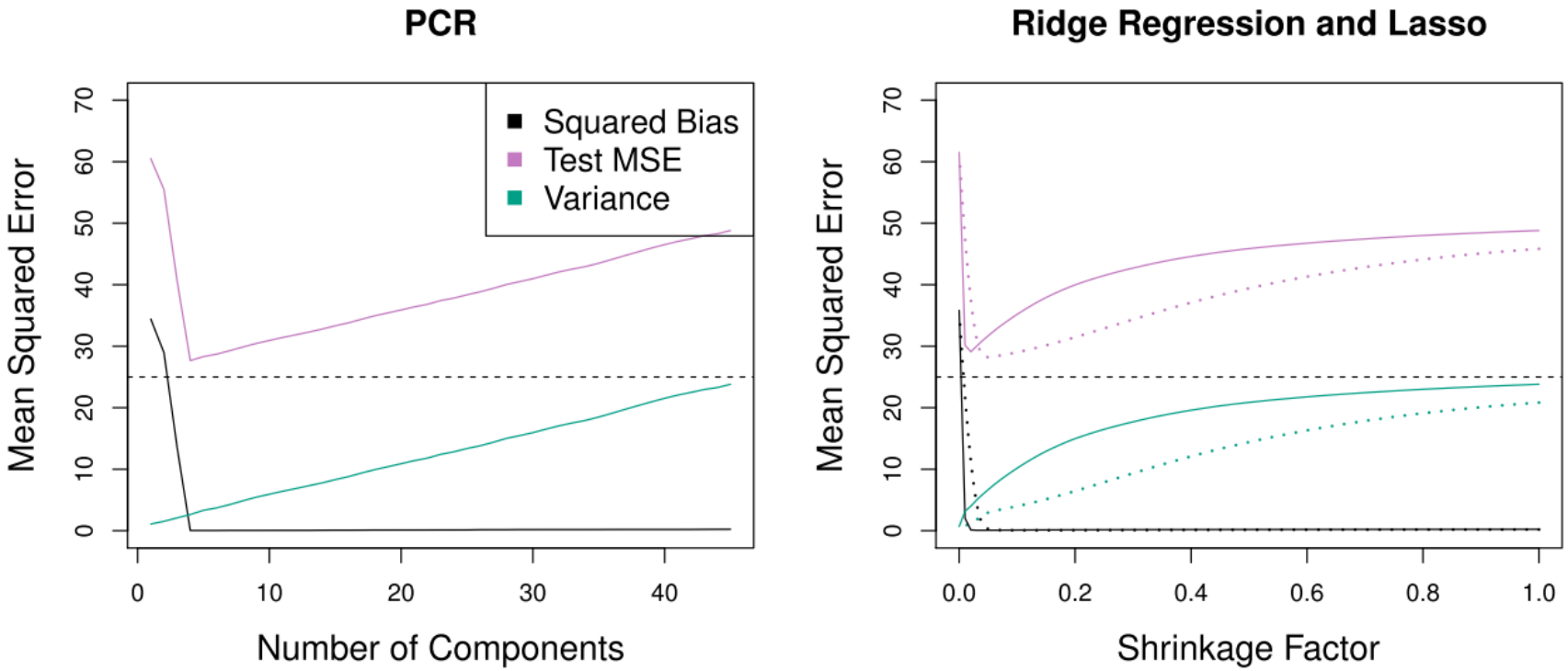
$n = 50$
$p = 45$ significant predictors



$n = 50$
$p = 2$ significant predictors

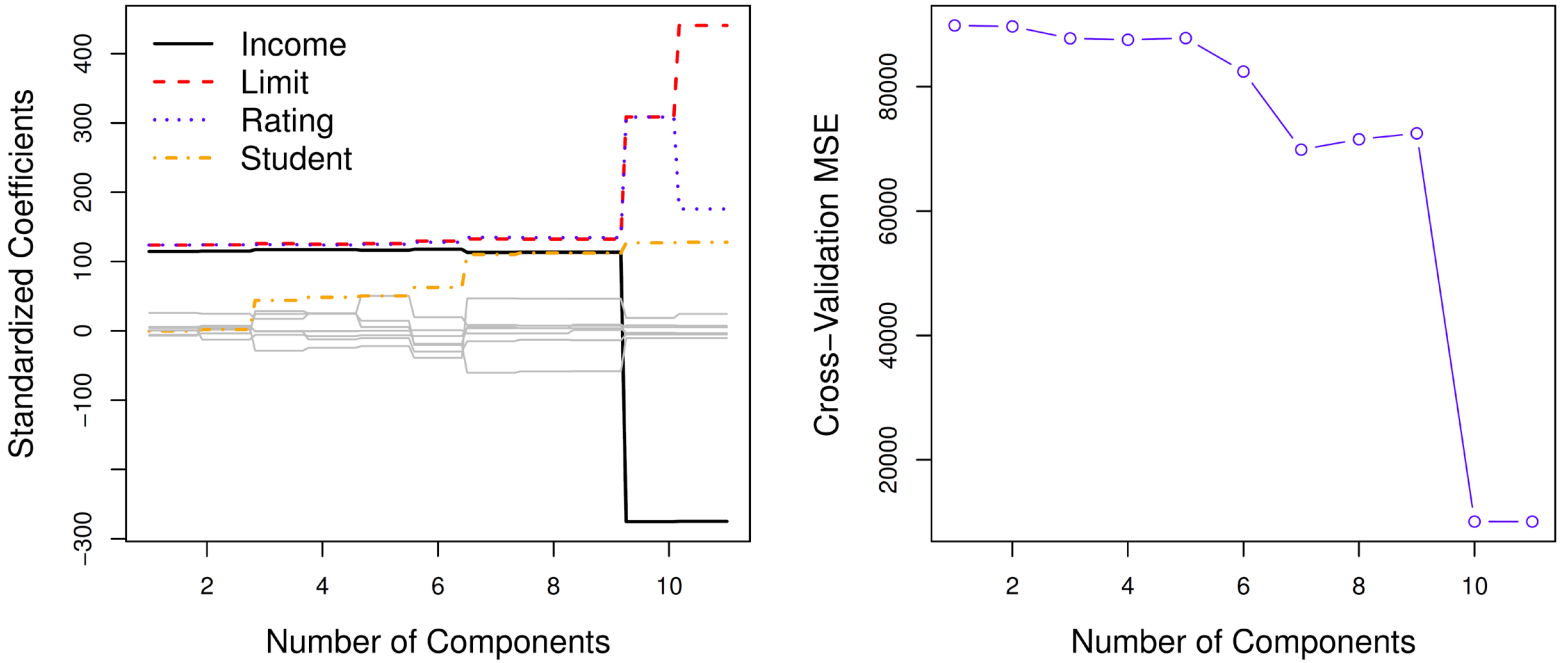
Application to Principal Components Regression

- PCR, ridge regression, and the lasso were applied to a simulated data set in which the first five principal components of X contain all the information about the response Y ($M = 5$). Lasso (solid) and ridge (dashed)



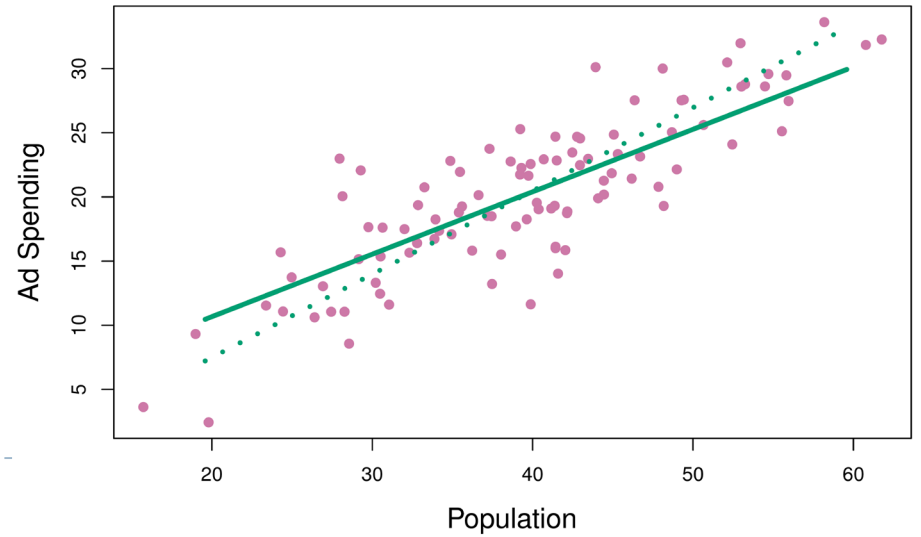
Choosing the number of directions M

- Left: PCR standardized coefficient estimates on the Credit data set for different values of M. Right: The 10-fold cross validation MSE obtained using PCR, as a function of M
  - Note that we also standardizing each predictor before PCR



Partial Least Squares

- PCR (dotted line) identifies linear combinations, or directions, that best represent the predictors $X_1, X_2, \ldots, X_p$
  - These directions are identified in an unsupervised way, since the response Y is not used to help determine the principal component directions
  - That is, the response does not supervise the identification of the principal components
  - Consequently, PCR suffers from a potentially serious drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response



Partial Least Squares: continued

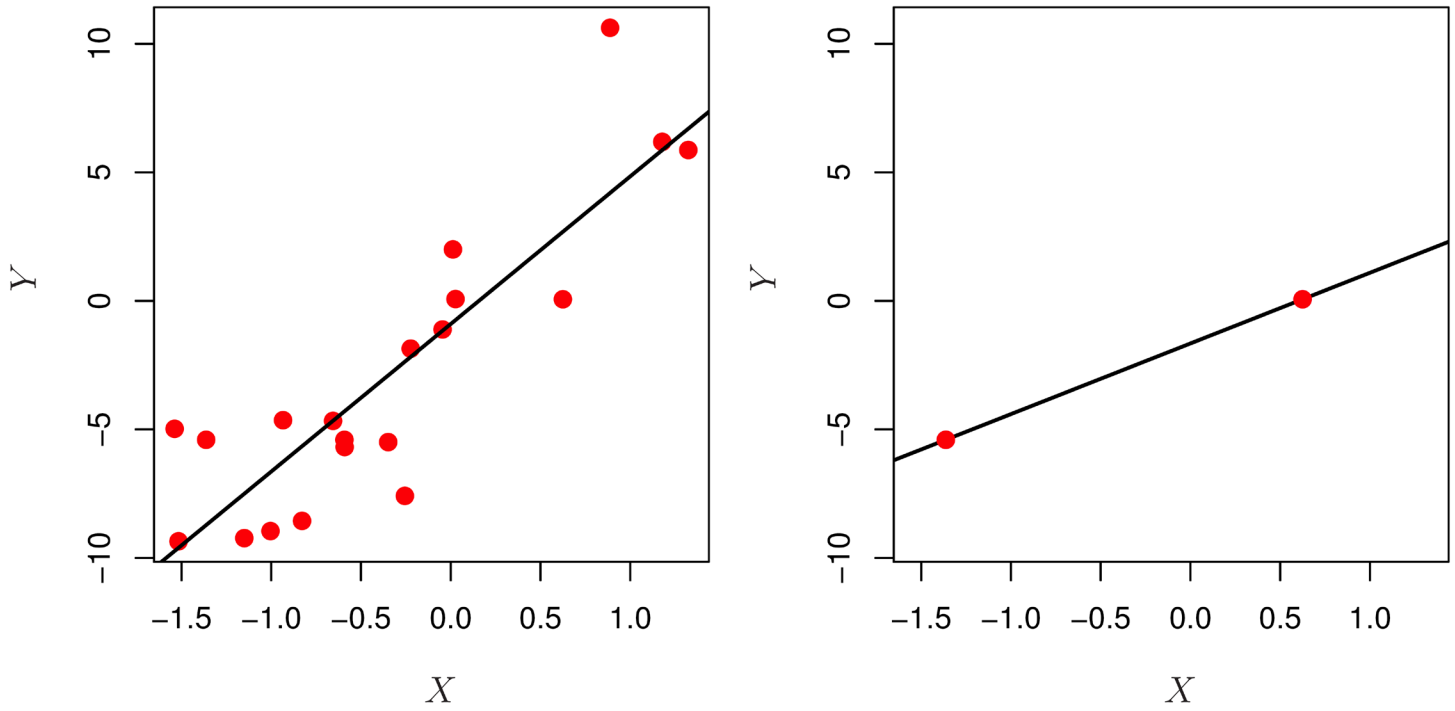
- Like PCR, PLS is a dimension reduction method, which first identifies a new set of features $Z_1, Z_2, \ldots, Z_M$ that are linear combinations of the original features, and then fits a linear model via OLS using these M new features
  - But unlike PCR, PLS identifies these new features in a supervised way - that is, it makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response
  - Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors

Details of Partial Least Squares

- After standardizing the p predictors, PLS computes the first direction Z_1 by setting Φ_{j1} equal to the coefficient from the simple linear regression of Y on X_j
  - One can show that this coefficient is proportional to the correlation between Y and X_j
  - Hence, in computing $Z_1 = \sum_{j=1}^{p} \Phi_{j1} X_j$ PLS places the highest weight on the variables that are most strongly related to the response
  - Subsequent directions are found by taking residuals and then repeating the above prescription (ESL 3.5)
  - More comparison about PLS to OLS

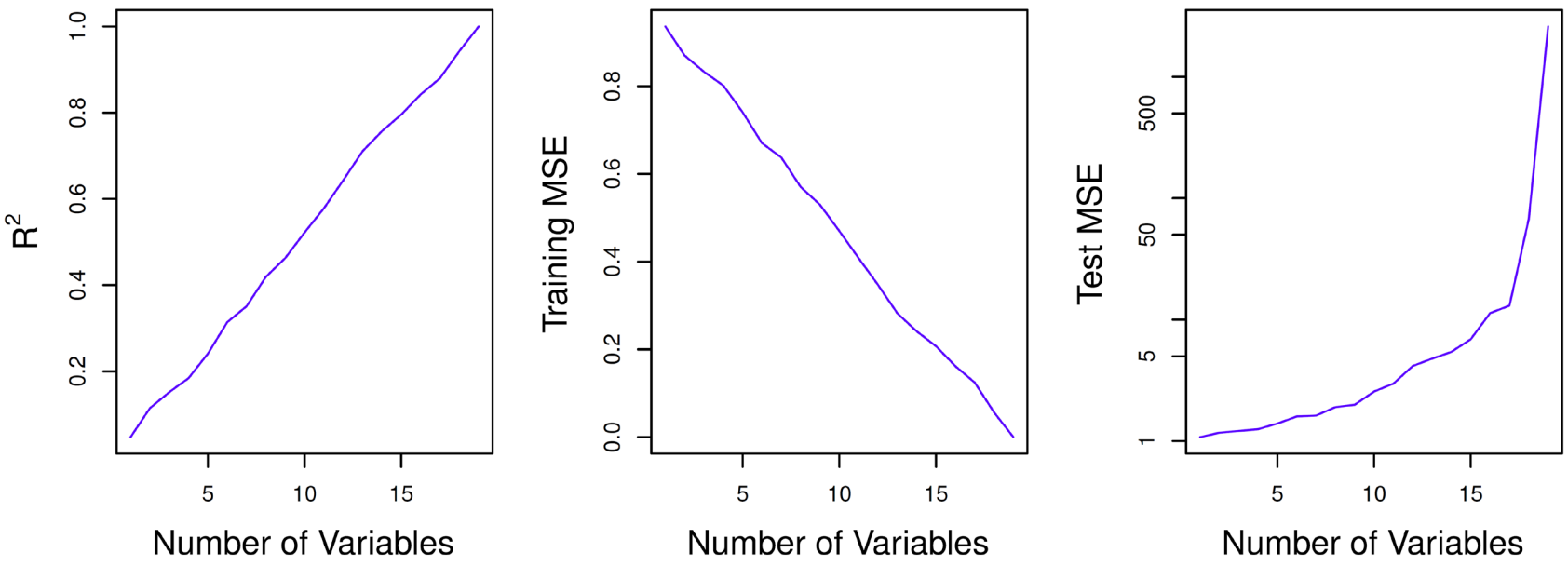
Consideration in high dimensions

- Left: Least squares regression in the low-dimensional setting. Right: Least squares regression with $n = 2$ observations and two parameters to be estimated (an intercept and a coefficient)

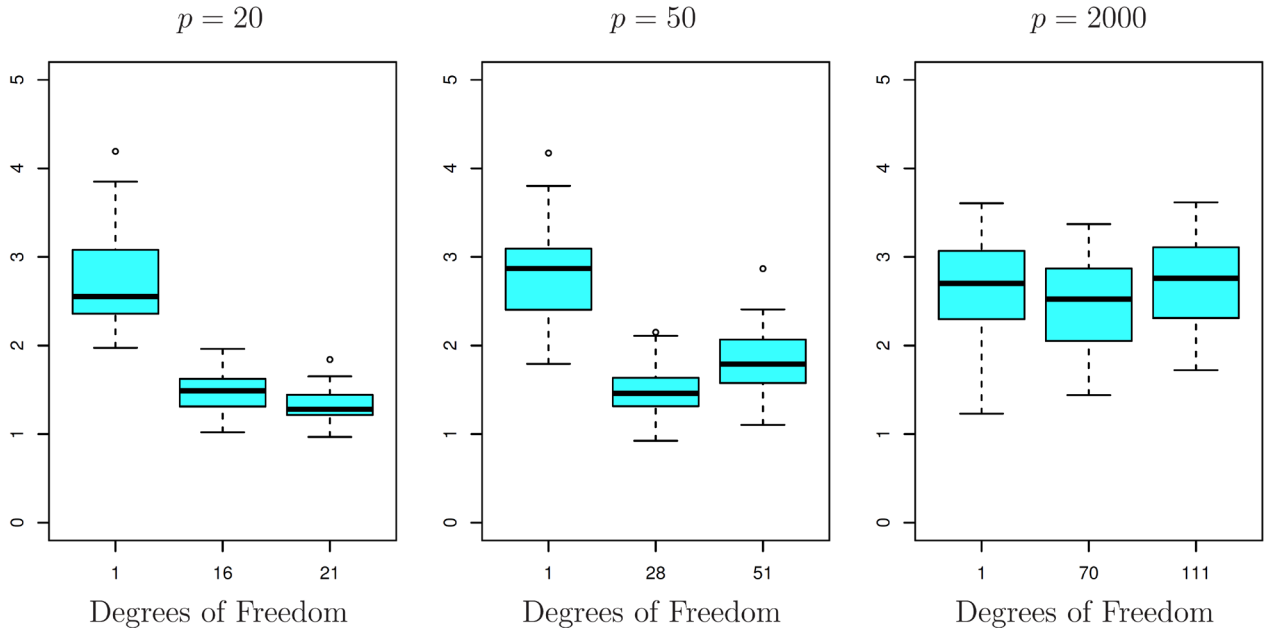


Consideration in high dimensions

- On a simulated example with $n = 20$ training observations, features that are completely unrelated to the outcome are added to the model
  - Left: The R^2 increases to 1 as more features are included. Center: The training set MSE decreases to 0 as more features are included. Right: The test set MSE increases as more features are included
  - This indicates the importance of always evaluating model performance on an independent test set



- The lasso was performed with $n = 100$ observations and three values of p, the number of features. Of the p features, 20 were associated with the response. The boxplots show the test MSEs that result using three different values of the tuning parameter



1. Regularization or shrinkage plays a key role in high-dimensional problems
2. Appropriate tuning parameter selection is crucial for good predictive performance
3. The test error tends to increase as the dimensionality of the problem (i.e. the number of features or predictors) increases, unless the additional features are truly associated with the response

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
- We should be careful when interpreting results in high dimensions

Appendix

Review of Covariance Matrix

- Let $x_1, \ldots, x_n$ be length-p observation vectors

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

- Without Loss Of Generality (WLOG), let their mean be length-p 0-vector
- Let the data matrix $X = (x_1, x_2, \ldots, x_n)$ be a p by n matrix
- The sample covariance matrix

$$S = XX^T/(n-1) = \sum_{i=1}^{n} x_i x_i^T/(n-1) = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T/(n-1)$$

Review of eigenvalue decomposition- Maximum variance formulation

- Find a direction vector $u_1 \in R^p$ and $u_1^T u_1 = 1$ such that the variance of the projected data is maximized

$$\frac{1}{n}\sum_{i=1}^{n}(u_1^T x_i - u_1^T \bar{x})^2 = u_1^T S u_1$$

  - To enforce the constraint, we introduce a Lagrange multiplier denoted by λ_1 and get the unconstrained maximization of

$$u_1^T S u_1 + \lambda_1(1 - u_1^T u_1) \text{ or maximize } \frac{u^T S u}{u^T u}$$

  - By setting the derivative with respect to u_1 equal to zero, we see that this quantity will have a stationary point when

$$S u_1 = \lambda_1 u_1$$

A is not a function of **x** **A** is symmetric	$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} =$	$2\mathbf{x}^\top \mathbf{A}$	$2\mathbf{A}\mathbf{x}$
u = **u**(**x**), **v** = **v**(**x**)	$\frac{\partial(\mathbf{u}\cdot\mathbf{v})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}^\top \mathbf{v}}{\partial \mathbf{x}} =$	$\mathbf{u}^\top \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^\top \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}, \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$ in numerator layout	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}\mathbf{u}$ $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}, \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$ in denominator layout

Review of eigenvalue decomposition- Maximum variance formulation

- u_1 must be an eigenvector of S, if we left-multiply by u_1^T we get

$$u_1^T S u_1 = \lambda_1$$

  - and so the variance will be a maximum when we set u_1 equal to the eigenvector having the largest eigenvalue λ_1. This eigenvector is known as the first principal component.

- We can define additional principal components in an incremental fashion by choosing each new direction to be that which maximizes the projected variance amongst all possible directions orthogonal to those already considered.
  - In a r-dimensional projection space, we now consider the optimal linear projection for which the variance of the projected data is maximized is defined by the r eigenvectors $u_1, \ldots, u_r$ of the data covariance matrix S corresponding to the r largest eigenvalues $\lambda_1, \ldots, \lambda_r$.

Principal Component Analysis (PCA) (1/2)

- If we collect eigenvectors and eigenvalues into matrix

$$S_{p\times p}U_{p\times p} = U_{p\times p}\Lambda_{p\times p}$$
$$S_{p\times p} = U_{p\times p}\Lambda_{p\times p}U^T_{p\times p}$$

- Note $X = USV^T$
  - Scores are $U^T X = SV^T$
- It is equivalent to Minimum error formulation

$$argmin_{U\epsilon\, O_{p,r}} \sum_{i=1}^{n} |(X_i - \bar{X}) - UU^T(X_i - \bar{X}))|_F^2$$

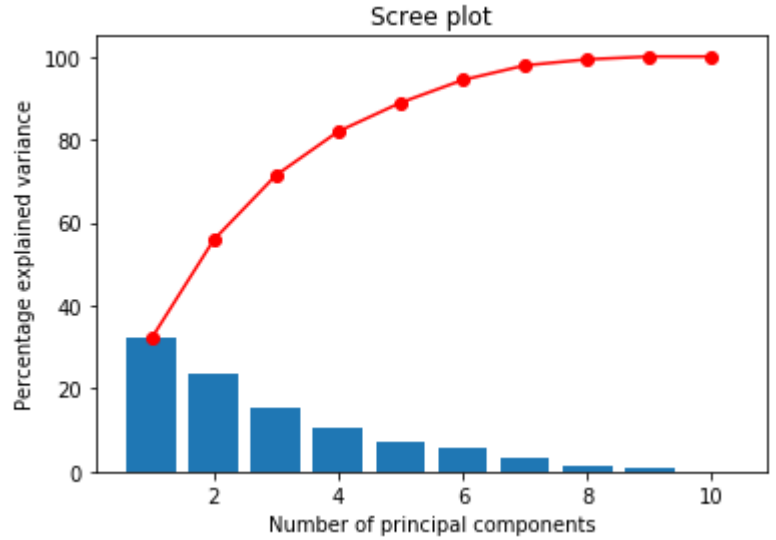
	Convention 1	Convention 2
U	Principal component Principal direction Loading	Principal axis Principal direction
$U^T X$	Principal component scores	Principal component

Principal Component Analysis (PCA) (2/2)

- Connection with SVD

$$S = \frac{XX^T}{n-1} = \frac{UDV^TVDU^T}{n-1} = U\frac{D^2}{n-1}U^T = U\Lambda U^T$$

- In practice, we will often scale data before PCA



- Whiten data matrix (identity covariance matrix)
  - $\Lambda^{-1/2}U^T\mathrm{X}$
- ZCA (Close to original data (often not reduce dimension))
  - $U\Lambda^{-1/2}U^T\mathrm{X}$

LAR and group Lasso in ESL

- Least Angle Regression
  - https://scikit-learn.org/stable/modules/linear_model.html#least-angle-regression
- Group Lasso
  - https://group-lasso.readthedocs.io/en/latest/