Ch=> 6.

Linear Model Selection & Regularization

In this chapter, we consider some approaches for extending the linear model framework.
In this chapter we will generalize the linear model in order to accomodate non-linear, but still additive relationships
→ Despite linear models simplicity, linear model has distinct advantages in terms of interpretability & often shows good prediction performance.
Hence in this chapter we consider we discuss some ways in which the simple linear model can be improved, by replacing ordinary least squares fitting with some alternative fitting procedures.

Why do we even need an alternatives to least squares?

- Prediction Accuracy: especially when $p \gg n$, to control the variance
- Model Interpretability: by removing irrelevant features that is, by setting the corresponding coefficient estimates to zero - we can find a model that is more easily interpreted
In this chapter we'll learn the approaches of feature selection.

Date :
Page :

Feature Selection can be done in 3 ways :

→ Subset Selection : identifies a subset of predictors that not only reduces least squares but also "variance & bias"

→ Shrinkage : fits a model with all "p" predictors but the estimated coefficients are shrunken to 0. This shrinkage (also known as regularization) reduces the variance & also does variable selection.

→ Dimension Reduction : We project our "p" predictors on a "M" dimensional subspace, where M < p.

(1.) Subset Selection : (Complexity : 2^p)

i.) Best Subset Selection :

1. Let (M_0) denote the null Model, which contains no predictors. This will simply predict the sample mean of the observations.
2. For $K = 1, 2, 3, \ldots, p$:
   (a) fit all (^pC_K) models that contains exactly K predictors
   (b) pick the best amongst the (^pC_K) models & call it M_K. Here best is defined as smallest RSS or largest R^2.
3. Select a single best model from among, $M_0, M_1, \ldots, M_p$ using cross validated predicted errors (AIC), BIC or adjusted R^2.

Disadvantages of Best Subset Selection :

(1) For computational reasons, best subset selection can not be applied with very large "p". Say $p = 10$, then total number of subsets ≈ 1000, but when $p = 40$, subsets = 1,000,000,000,000

& that is computationally very expensive.

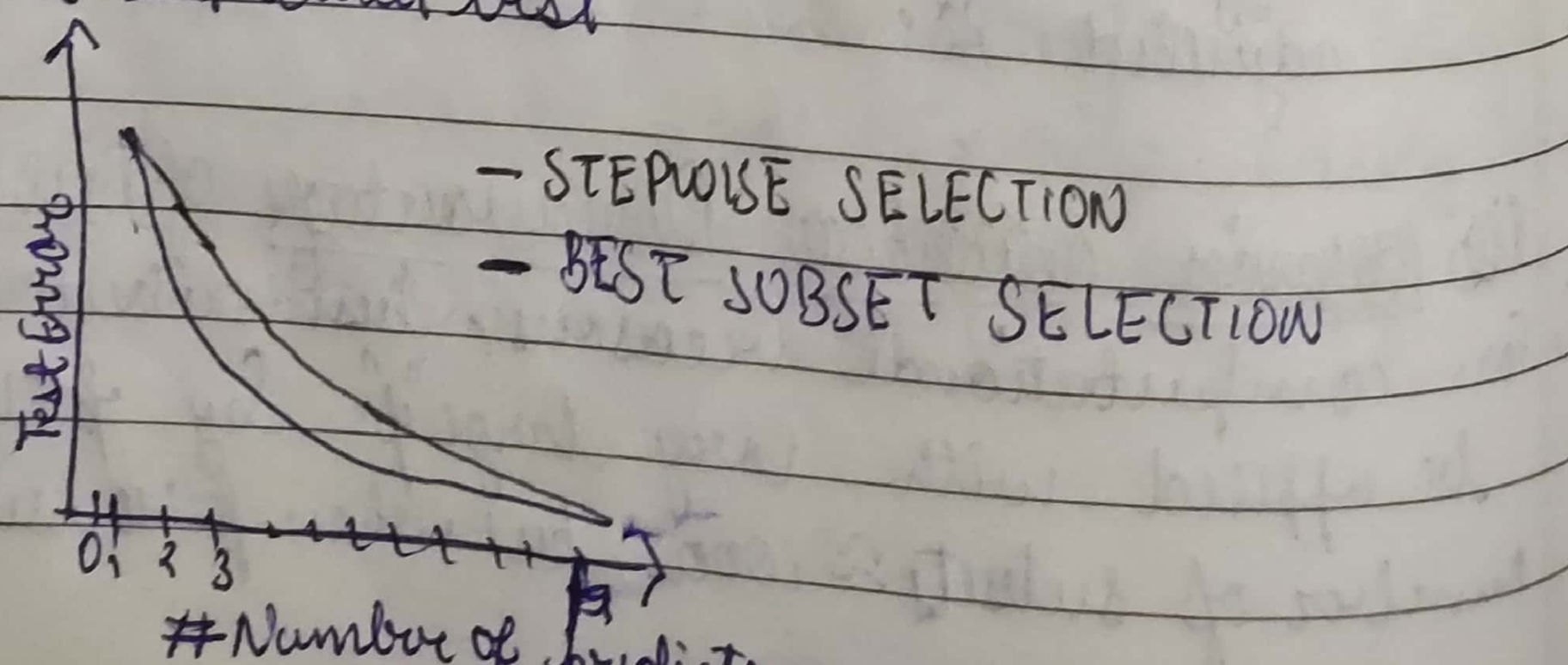
(2) It will also suffer from statistical problems when "p" is large. The bigger the search data the bigger the chances of finding a model that works good on training data but might not have any power on future data.
Thus an enormous search space can lead to overfitting.

ii) STEPWISE SELECTION: (Complexity: p^2)
For the reasons mentioned above stepwise methods would work better.

(1) Forward Stepwise Selection:
→ This begins with a model containing no predictors and adds predictors to the model one at a time, until all predictors are in model. At each step the variable that gives the best improvement to the fit is added to the model.

→ STEPS:
1. Let M_0 denote a null Model, which contains no predictors
2. For $K = 0, 1, 2, \ldots, p-1$:
   2.1 Consider all "$p-K$" models that augments the predictors in M_K with one additional predictor.
   2.2. Choose the best amongst all the models available
3. Select a single best model from M_0 to M_p.
→ Computational advantage over best subset selection.
→ But this is not guaranteed that the best model selected will be the actual best

→



→ Why does best subset selection provides lower test error? Reason being there might be some positive correlation between features which cannot be found in best subset section.

→ Why does the graph start & end at the same point? because they both start with 0 predictors & ends with all p predictors.

② Backward Stepwise Selection:

→ Backward stepwise selection provides an efficient alternative to best subset selection like forward stepwise selection.

→ However, it starts with all "p" predictors & removes one at a time.

→ STEPS:

1. Let Mp denote the full model with all "p" predictors
2. For k = p, p-1, p-2 ..... , 1:
   - 2.1 Consider all k models with one removed predictor.
   - 2.2 Choose the Best model with least RSS or highest R^2.
3. Select a single best model from $M_0, \ldots, M_p$

CHOOSING THE OPTIMAL MODEL:

→ The chosen model must have lowest RSS & highest R^2.

→ we choose to with a model with low test error not train error.

→ Therefore, RSS & R^2 are not good predictor for best model.

ESTIMATING THE TEST ERROR:

→ We can indirectly estimate test error by making an adjustment to the training error to account for the overfitting.

→ We can directly estimate test error using a validation set approach or the cross validation approach.

Adjusting LOSS FUNCTION: (Cp, AIC, BIC)

→ Mallow's Cp:

$$Cp = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

where, d = # of parameters

$\hat{\sigma}^2$ = estimate of variance of error term.

→ AIC: (AKaike Information criteria)

The AIC is defined for a large number of models fit by maximum liklihood.

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

where, L is the maximized value of the likelihood model

→ BIC: (Bayesian Information Criteria)

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

→ Adjusted $-R^2$:

$$1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Date :
Page :

Validation & Cross Validation.

→ We can use validation or cross validation for selecting the model with least test error.

→ This has an advantage over AIC, BIC, C_p, adjusted R^2 that it provides direct estimate of test error.

② Shrinkage Methods:

The subset selection uses a subset of predictor to fit a linear model with least squares, shrinkage methods on the other hand fits the model with all "p" predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently shrinks the coefficient estimates towards zero.

i) Ridge Regression:

Recall, that least squares is given by:

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

On the other hand ridge regression adds a penalty to RSS, & β_R tries to minimize least square for the equation below:

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

$$= RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

where, $\lambda \geq 0$ is a tuning parameter, i.e. a hyperparameter

→ as with least squares, ridge regression seeks coefficient estimates that fit the data well by making RSS small.

→ however, second term in ridge, i.e. $\lambda \sum \beta_j^2$, called a shrinkage penalty is small when $\beta_1, \beta_2, \beta_3 \cdots \beta_p$ are close to zero; so it has the effect of shrinking β_j to 0.

→ so, this will automatically shrink less important parameters towards zero.

→ λ here serves as tuning parameter, higher the lambda lower the value of coefficients.

→ The penalty part is also called l_2-norm & is given by : $\|\beta\|_2$

$$\|\beta\|_2 = \sqrt{\sum \beta_j^2}$$

Scaling the predictors in ridge regression is pretty important, why?

→ Multiplying x_j, with a constant "c", simply leads a scaling of least square coefficient estimates by a factor of $1/c$.

→ so, because of the scaling in ridge regression coefficient estimate will lead to change in the output of the loss in because of the presence of the term $\sum \beta_j^2$ in the penalty.

→ Scaling wasn't important in ordinary least squares but is important here because of the presence of penalty terms.

∴ it is best to apply ridge regression after standardization

Date :
Page :

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

iii) The LASSO :

→ Ridge Regression has one obvious disadvantage :
Unlike subset selection, which selects "s" parameters out of "p", ridge might reduce the coefficients close to zero but not exactly zero.

→ The Lasso is a relatively recent alternative to ridge that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_L$ minimize the quantity

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

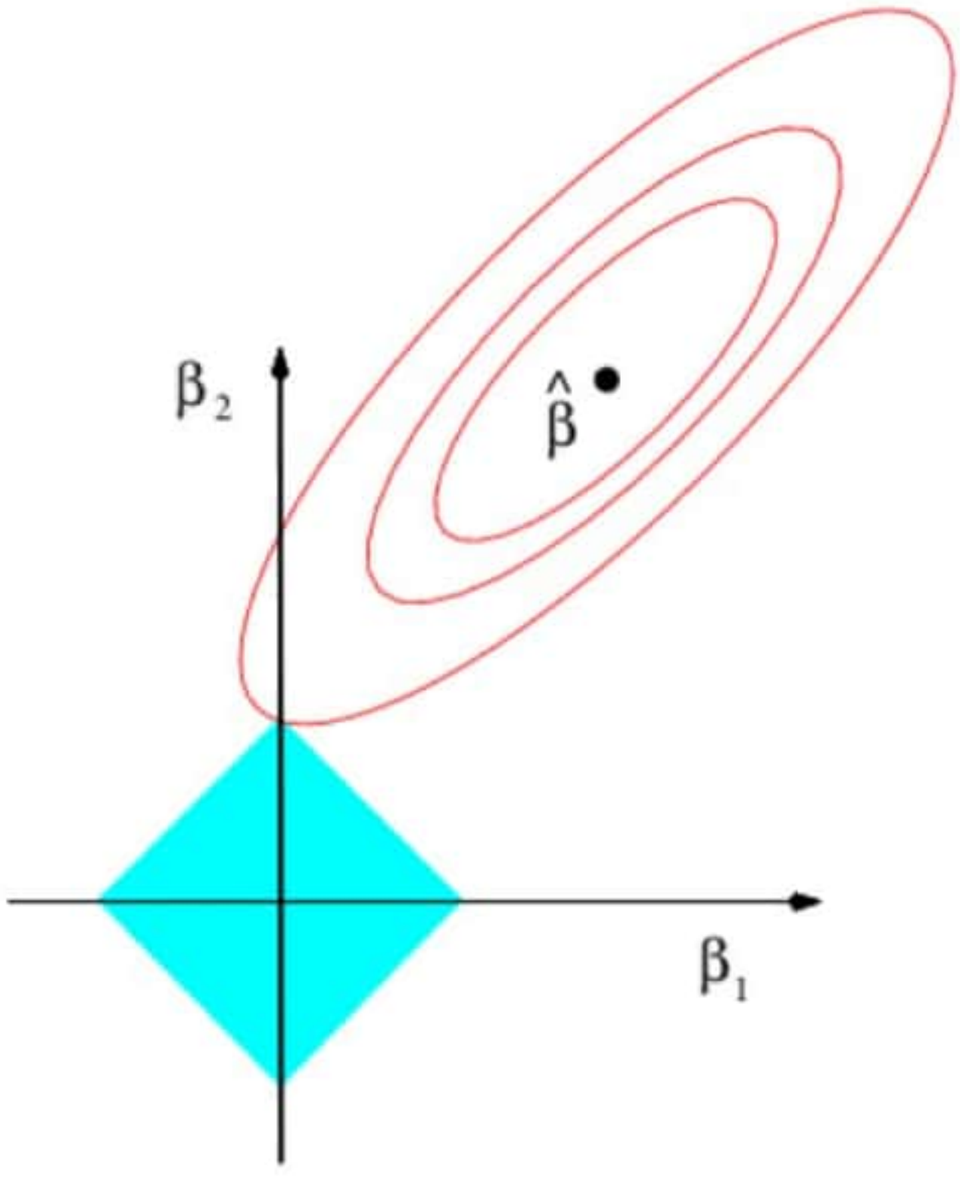
→ In statistical parlance, the lasso uses an l_1 penalty ($||\beta_j||_1$) instead of l_2 penalty

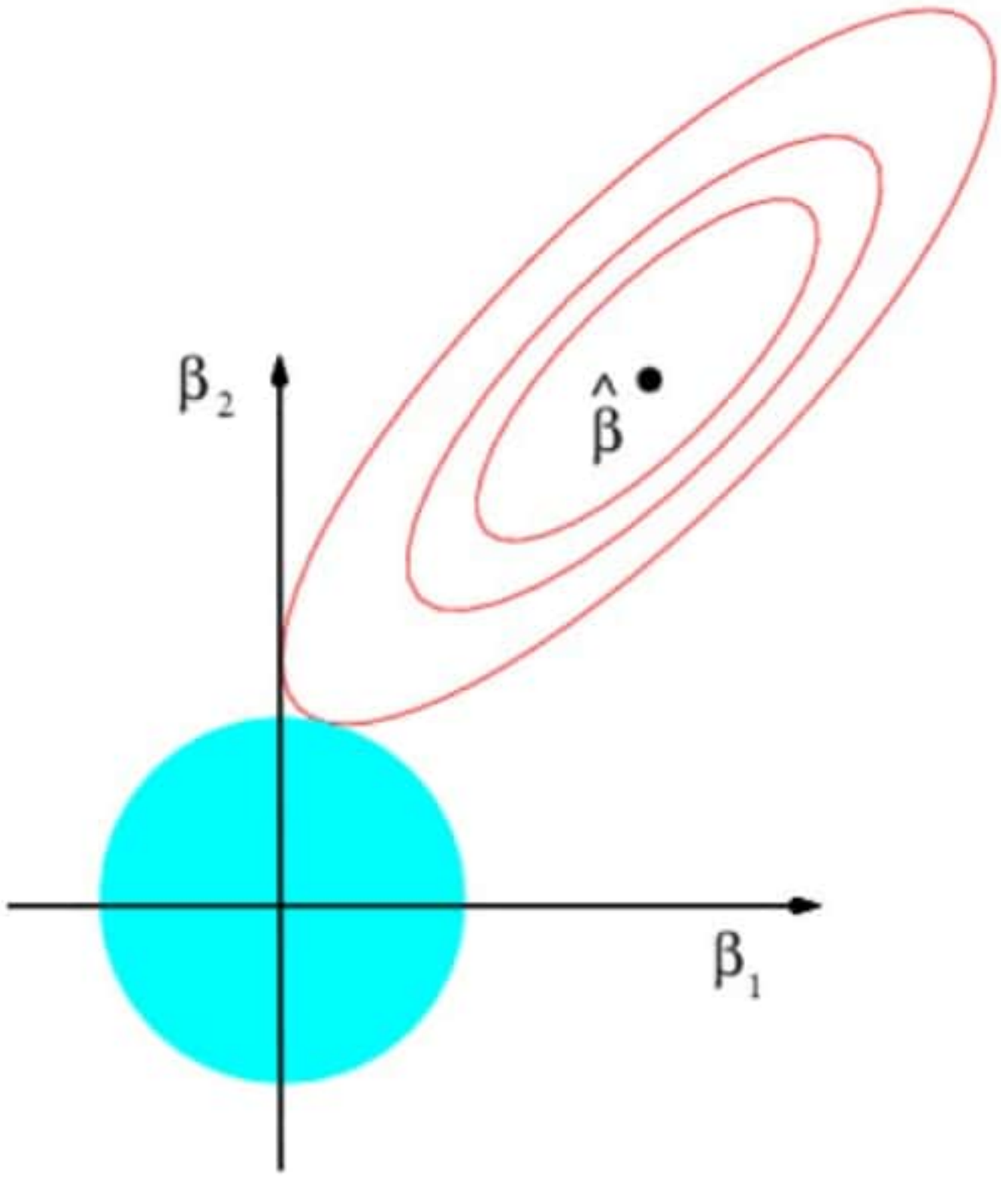
→ However, both ridge & lasso shrinks coefficient estimates towards zero, but lasso has an effect of forcing some of the coefficents towards zero.

→ Hence like subset selection lasso does variable selection.

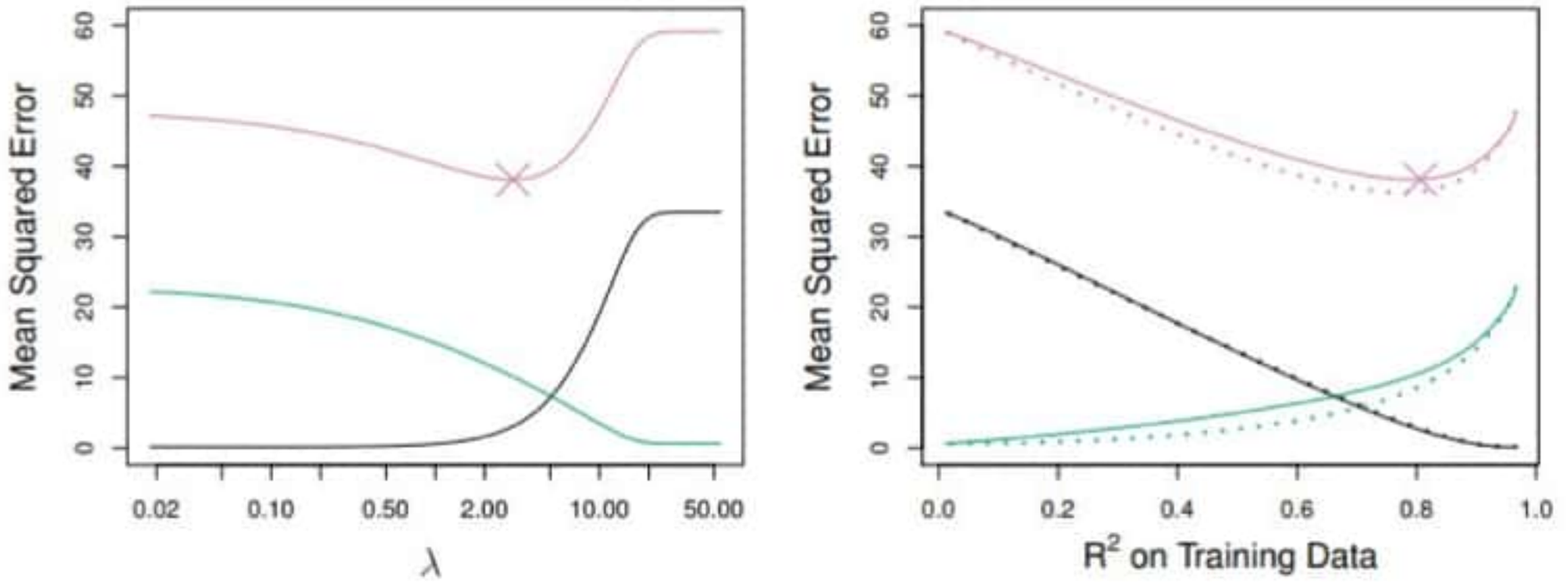
→ We say lasso is a sparse model while ridge is a dense model.

The Lasso Picture



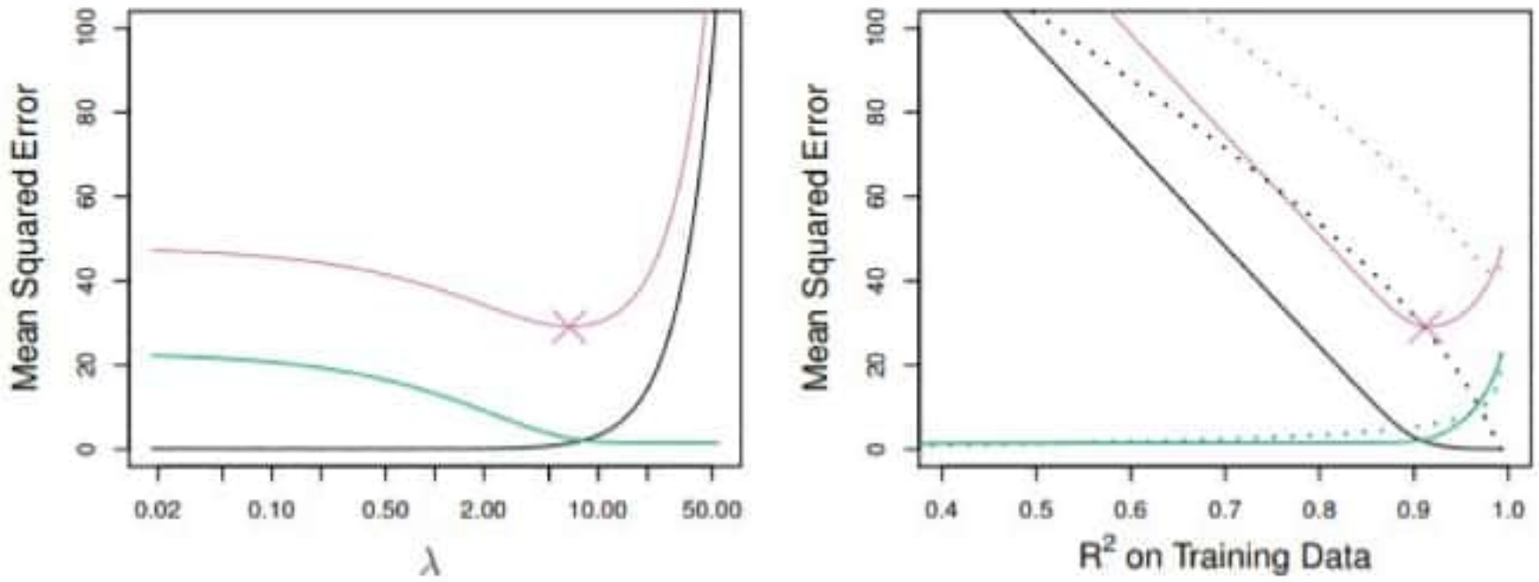


Comparing the Lasso and Ridge Regression



Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on simulated data set of Slide 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.

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 38, 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. The crosses in both plots indicate the lasso model for which the MSE is smallest.