Shrinkage methods & Regularization

Pham Mai Tam

15th June, 2022

Outline

1. Shrinkage Methods
   1.1 Ridge Regression
   1.2 The Lasso
      - ☞ Comparing the Lasso and Ridge Regression
      - ☞ Bayesian Interpretation for Ridge Regression and the Lasso
2. Selecting the Tuning Parameter (Lamda, λ)

Shrinkage Methods

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 $\beta_0, \beta_1, \ldots, \beta_p$ 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 $\hat{\beta}^R$ 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**.

→ Aim of λ: penalizing large values of parameters in order to prevent over-fitting to your training data

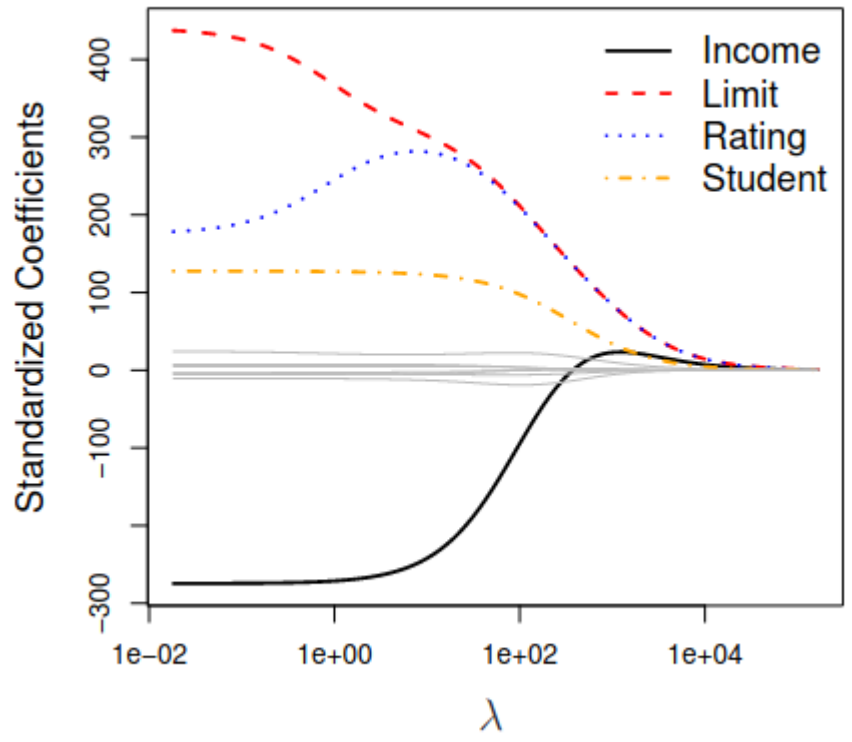
Features

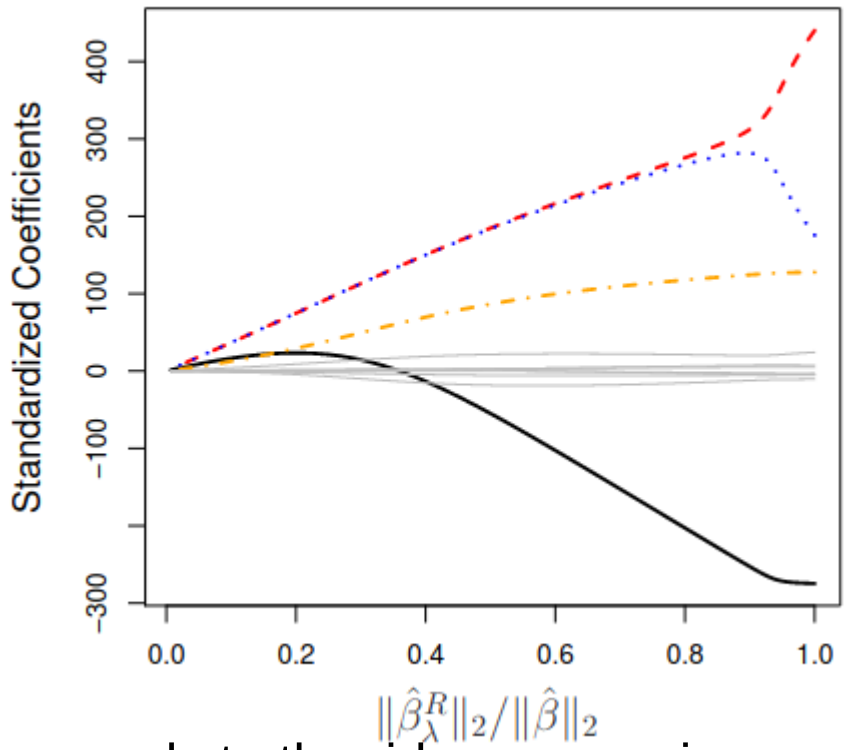
- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
- However, the second term, $\lambda \sum_j \beta_j^2$, called a **shrinkage penalty**, is small when $\beta_1, \ldots, \beta_p$ are close to zero, and so it has the effect of shrinking the estimates of β_j towards **zero**.
- The tuning parameter λ serves to control the relative impact on the regression coefficient estimates.
- $\lambda = 0$, penalty has no effect → as RSS only
- $\lambda \to \infty$, impact of shrinkage penalty **grow** → ridge regression coefficient estimates will approach **zero**.

→ Unlike least squares (generates only 1 set of coefficient estimate), ridge regression will produce a different set of coefficient estimates for each value of λ, called $\hat{\beta}_\lambda^R$.

→ Selecting a good value for λ is critical; **cross-validation** is used for this.

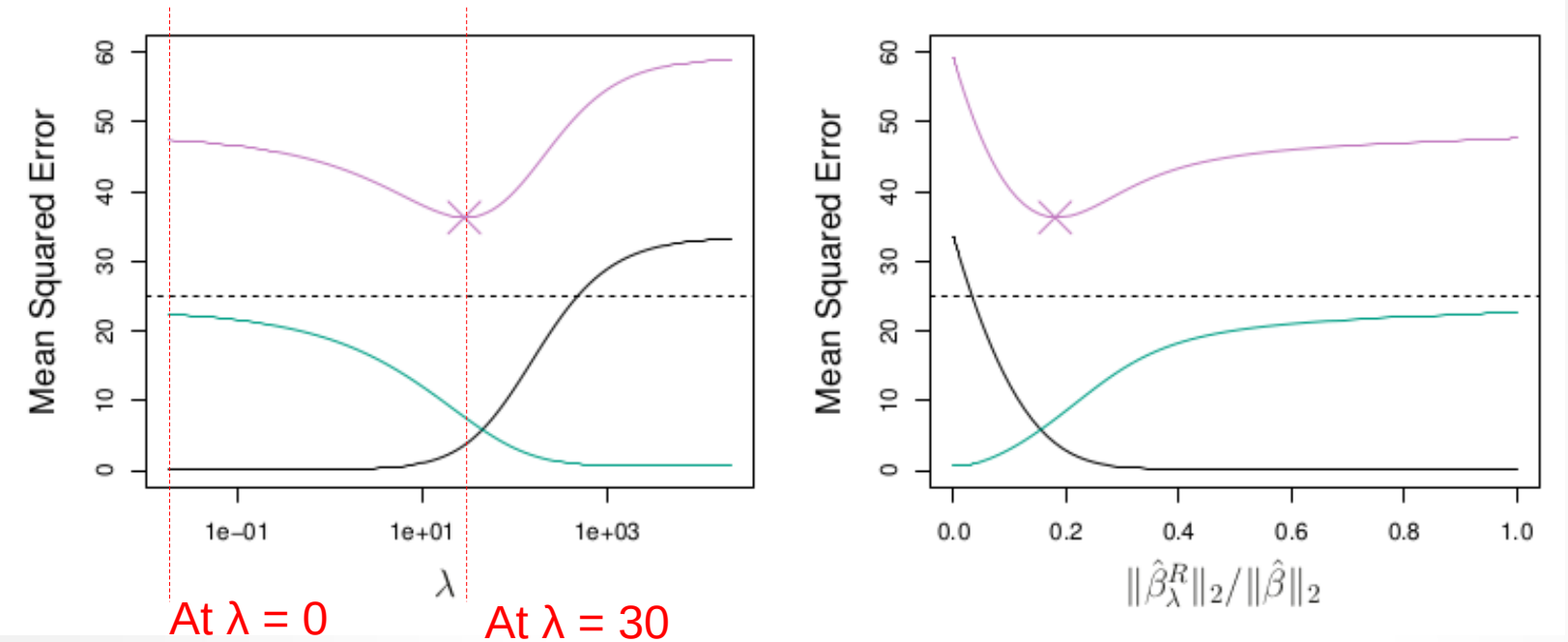
Lambda and coefficients





- 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 λ 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 ℓ_2 norm (pronounced “ell 2”) of a vector, and is defined as $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j{}^2}$

Why does ridge regression improve over least squares? → bias variance trade-off



Purple: test MSE
Black: squared bias
Green: variance

- Simulated data with n = 50 (sample size), p = 45 predictors, all having nonzero coefficients.
- The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate λ=30 at which the MSE is smallest.

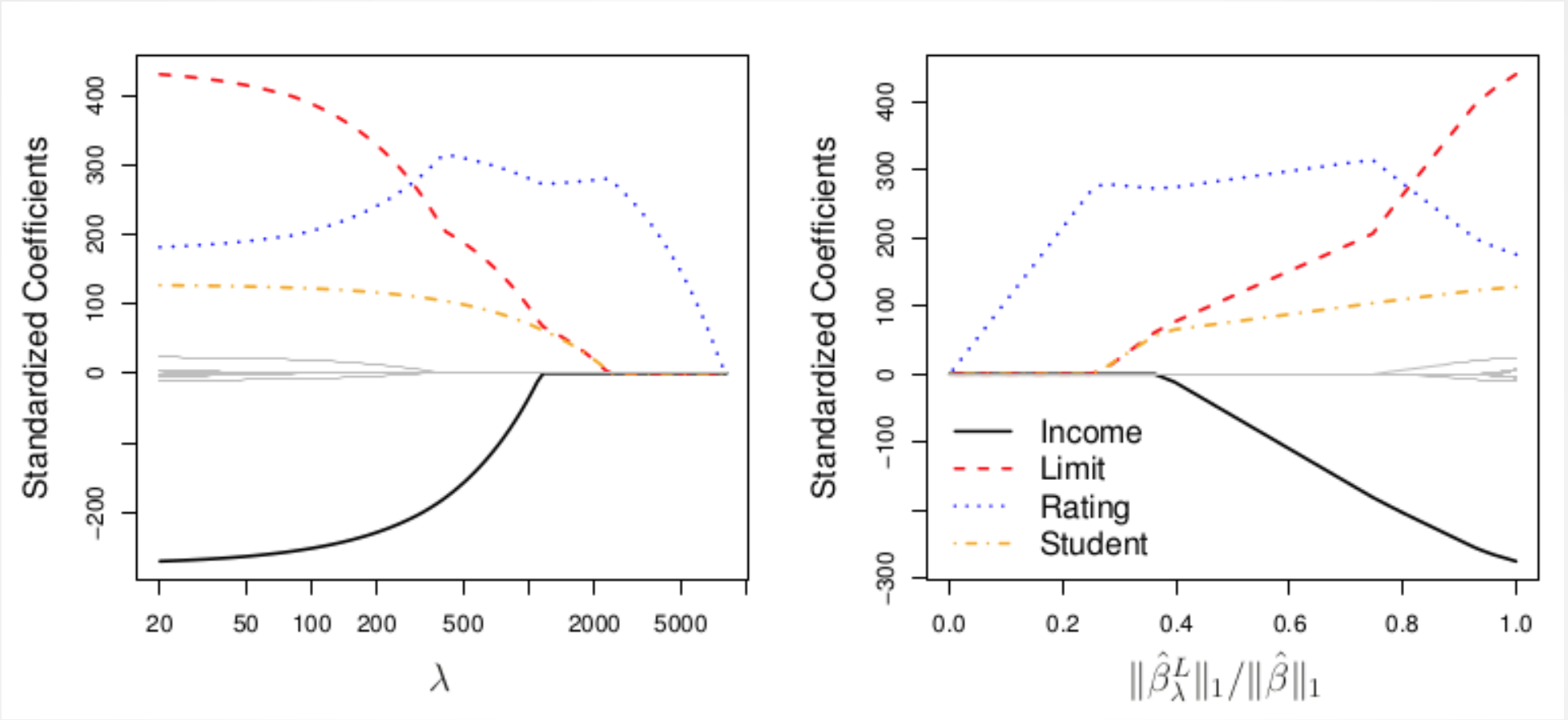
The Lasso

- 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

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

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

The Lasso


Standardized Coefficients
400
300
200
100
0
−200
20
50
100
200
500
2000
5000
λ
Standardized Coefficients
400
300
200
100
0
−100
−300
0.0
0.2
0.4
0.6
0.8
1.0
Income
Limit
Rating
Student
$\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$

Variable Selection Property of the Lasso

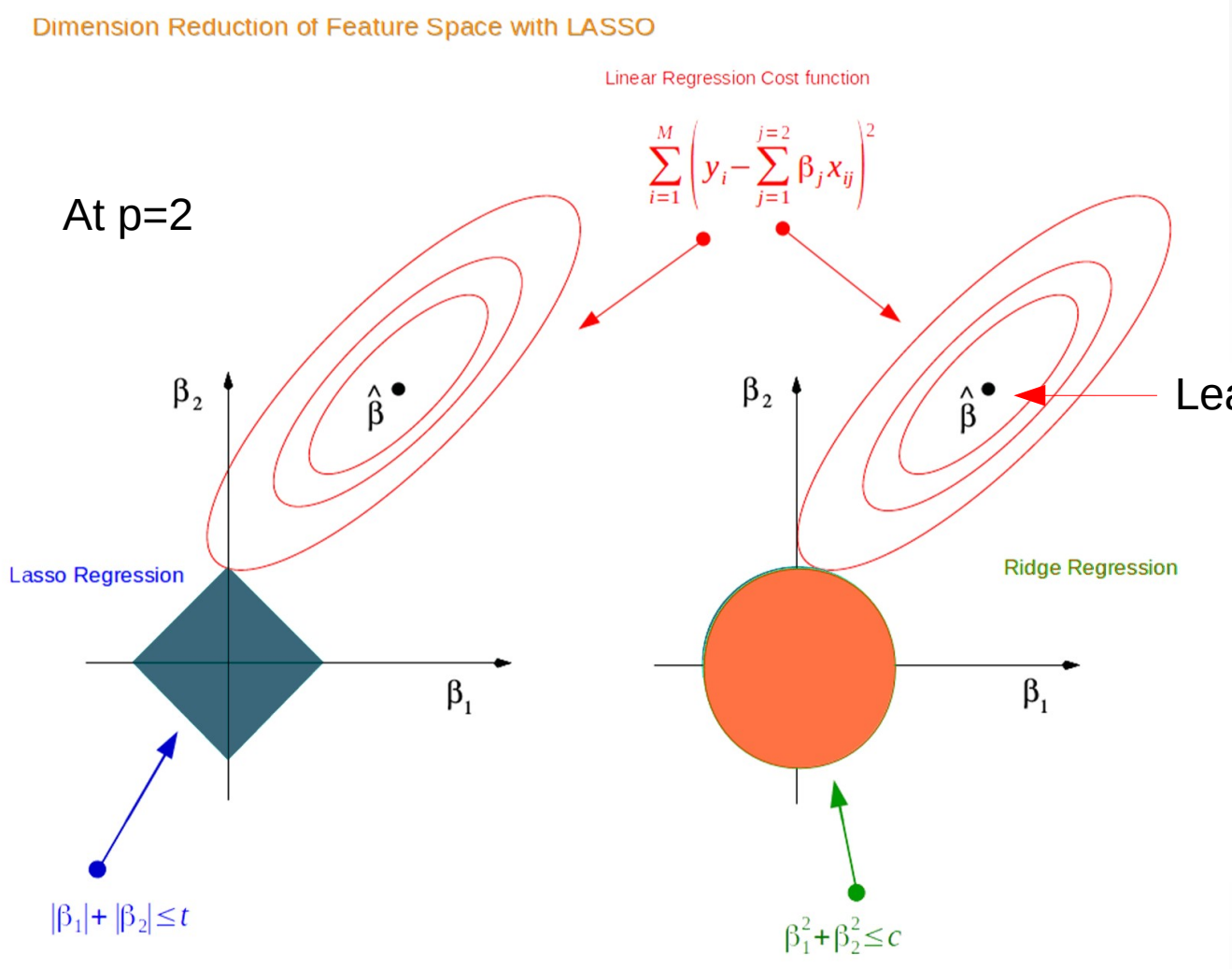
Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?
Larange multiplier show that the lasso and ridge regression coefficient estimates solve the problems

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

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

Lagrange multiplier (λ)

Lasso Picture



Lasso: coefficients can be exactly zero
Ridge regression: coefficients towards zero, never be exactly zero

Least squares solution

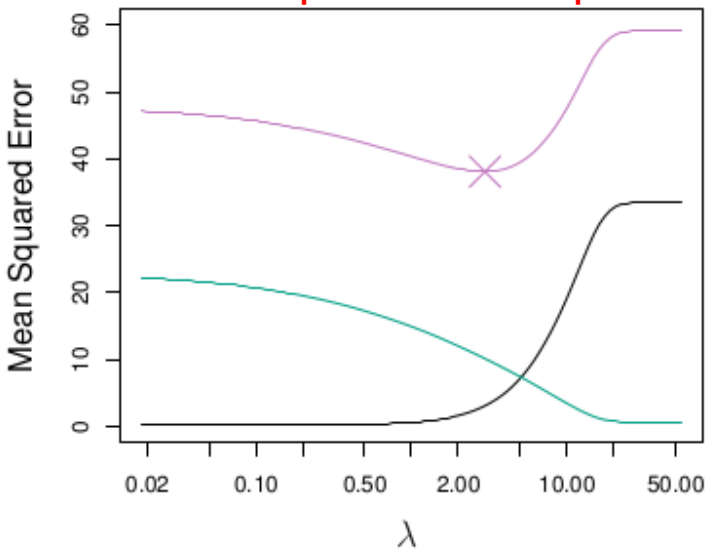
If s is sufficiently large (λ = 0), constrain region will contain $\hat{\beta}$,

$$\hat{\beta}^R = \hat{\beta}_\lambda^L = \hat{\beta}$$

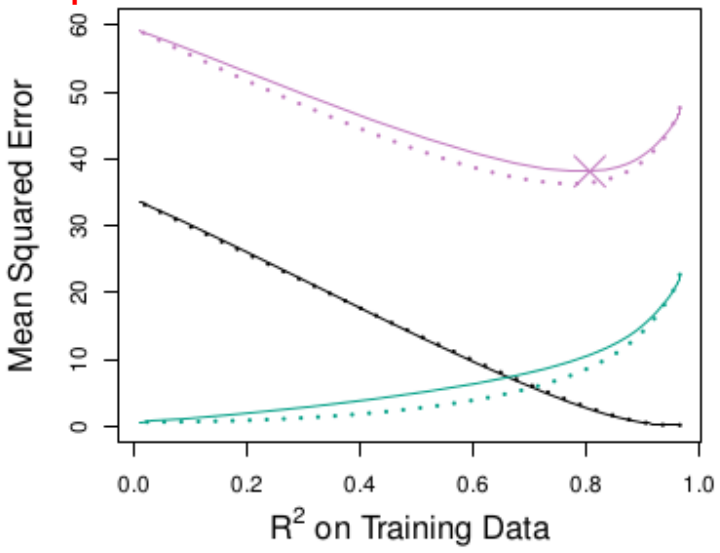
→ Indicate that the lasso and ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint region.

When to use the Lasso and Ridge Regression

When most predictors impact the response



the lasso on simulated data set

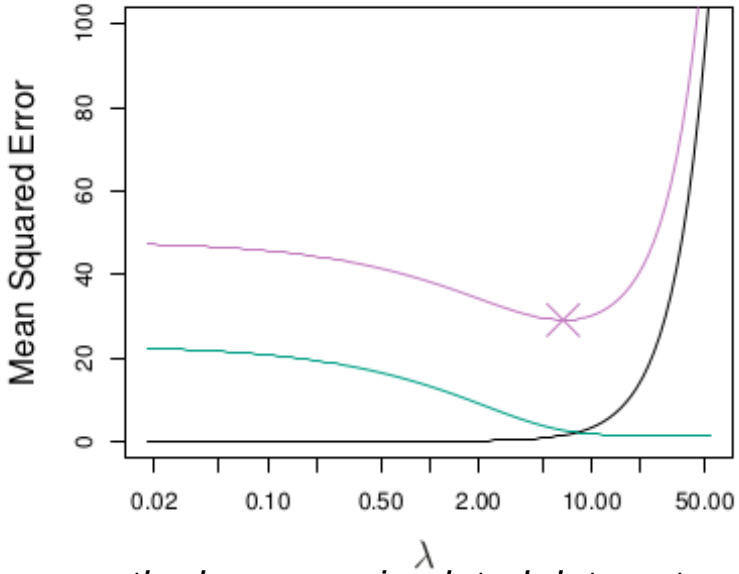


lasso (solid) and ridge (dashed)

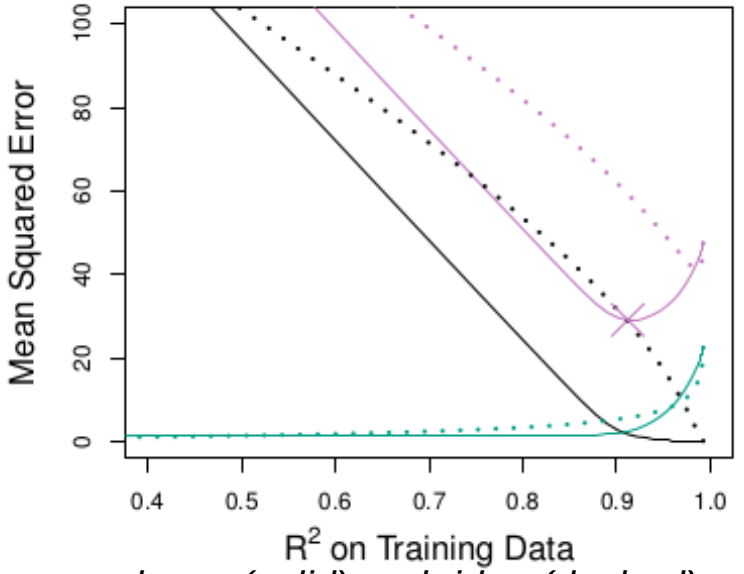
Purple: test MSE
Black: squared bias
Green: variance

→ **Ridge regression performs better the lasso** if there are many large parameters of about the same value

When when only a few predictors actually influence the response



the lasso on simulated data set



lasso (solid) and ridge (dashed)

→ **The lasso performs better ridge regression** if there are a small number of significant parameters and the others are close to zero

→ These 2 cases showed that neither ridge regression nor the lasso will universally dominate the other

Bayesian Inference for Ridge Regression and the Lasso

*'Instead of thinking about the line **minimize the cost**, think about it as **maximizing the likelihood** of the observed data'*

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

is the predicted outcome for the ith observation. We can write the actual value of the ith observation as:

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i$$

To get the likelihood of our data given our assumptions about how it was generated, we must get the probability of each data point y and multiply them together.

$$\text{likelihood} = p(y_1|x_1, \theta) * p(y_2|x_2, \theta) \ldots * p(y_n|x_n, \theta)$$

We want to find values of θ that maximize this result.

Assume ϵ is normally distributed

Back to calculating the probabilities of our observed y values... if each y_i is $\theta^T x_i + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, we can also say that

$$y_i \sim \mathcal{N}(\theta^T x_i, \sigma^2)$$

Recall that:

$X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(0, \sigma^2)$ then $X = \mu + Y$

Bayesian Inference for Ridge Regression and the Lasso

We want to maximize the likelihood of our data with respect to our parameters, θ. Here's that likelihood shown as the product of normal densities:

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\theta^T x_i)^2}{2\sigma^2}}$$

With a little mathematical shunting (and remembering that $e^x e^y = e^{x+y}$) we can see that this is equivalent to:

$$\sigma^2 2\pi^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i-\theta^T x_i)^2}$$

→ maximize likelihood is minimize RSS

MAP estimates and Ridge Regression

Bayes' theory

$$\text{posterior} = \frac{\text{likelihood x prior}}{\text{marginal likelihood}}$$

or

$$\text{posterior} \propto \text{likelihood x prior}$$

Recall cost function of Ridge Regression

$$\arg\min_{\theta} \sum_1^n (y_i - \hat{y}_i)^2 + \lambda \sum_1^p \theta^2$$

Now let's see how we can arrive at this same solution from the Bayesian method. Recall that the posterior distribution for the weights is proportional to the likelihood times the prior. Using the likelihood we figured out before and a Gaussian prior with a mean of 0 and a variance of τ^2, this becomes:

$$\sigma^2 2\pi^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2} \times \tau^2 2\pi^{-\frac{p}{2}} e^{-\frac{1}{2\tau^2}\sum_1^p \theta^2}$$

with the likelihood in blue and the prior in green. With some slight rearranging we get

$$e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2} \times \sigma^2 2\pi^{-\frac{n}{2}} \times \tau^2 2\pi^{-\frac{p}{2}}$$

Given that we are looking to maximize this with respect to the coefficients, we can ignore the terms that don't depend on the coefficients:

$$\arg\max_\theta \; e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2}$$

And remember that we like to work with the log-likelihood instead of the likelihood, so now we have

$$\arg\max_\theta \; -\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2$$

Multiplying by $2\sigma^2$ and pulling out the -1 we get:

$$\arg\max_\theta \; -1(\sum_1^n (y_i - \hat{y}_i)^2 + \frac{\sigma^2}{\tau^2}\sum_1^p \theta^2)$$

And since maximizing $-x$ is equivalent to minimizing x, this is equivalent to:

$$\arg\min_\theta \sum_1^n (y_i - \hat{y}_i)^2 + \frac{\sigma^2}{\tau^2}\sum_1^p \theta^2$$

And this is exactly what we have above for a regularized cost function, only with $\frac{\sigma^2}{\tau^2}$ instead of λ.

Relationship of posterior distribution and Ridge regression/Lasso

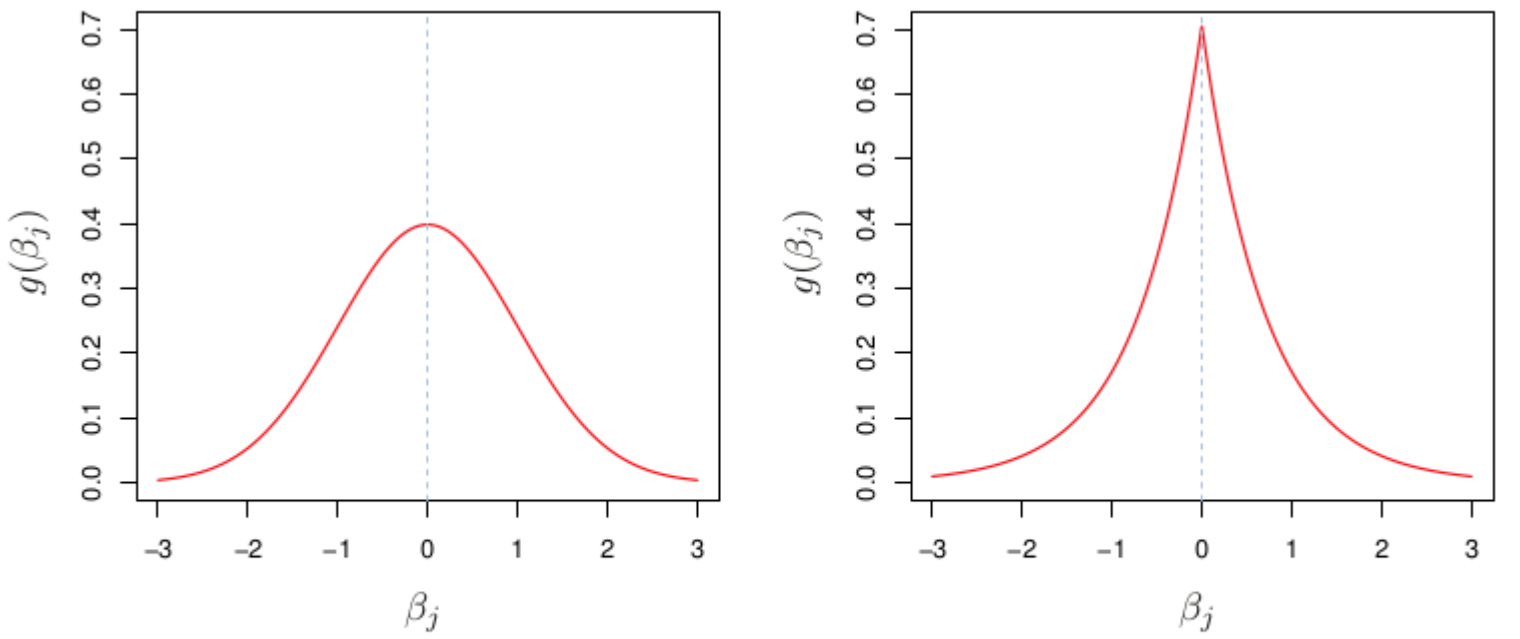


FIGURE 6.11. Left: *Ridge regression is the posterior mode for β under a Gaussian prior.* Right: *The lasso is the posterior mode for β under a double-exponential prior.*

- If a prior follow **Gaussian distribution** with mean zero and standard deviation a function of λ, then it follows that the posterior mode for β is **ridge regression** solution
- If a prior follow **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

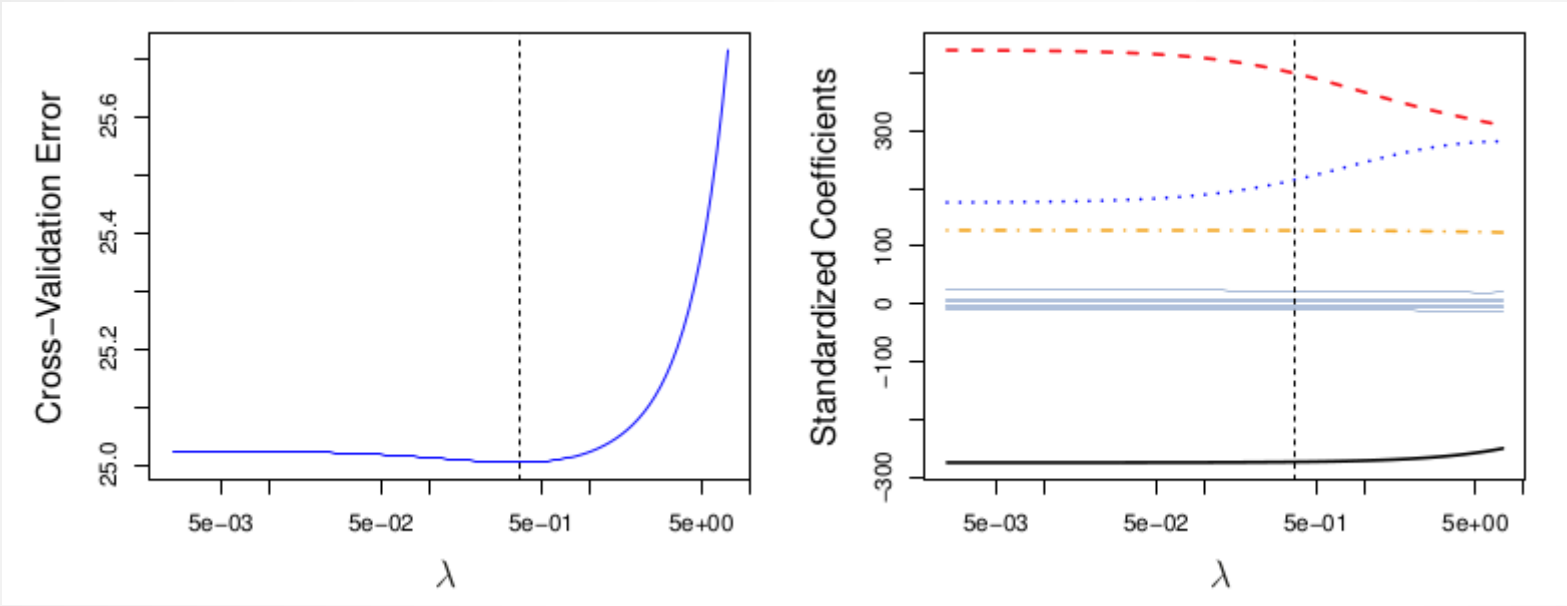
Selecting Tuning Parameter for Ridge Regression and Lasso

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration is best (lowest test MSE).
- That is, we require a method selecting a value for the tuning parameter λ or equivalently, the value of the constraint s.

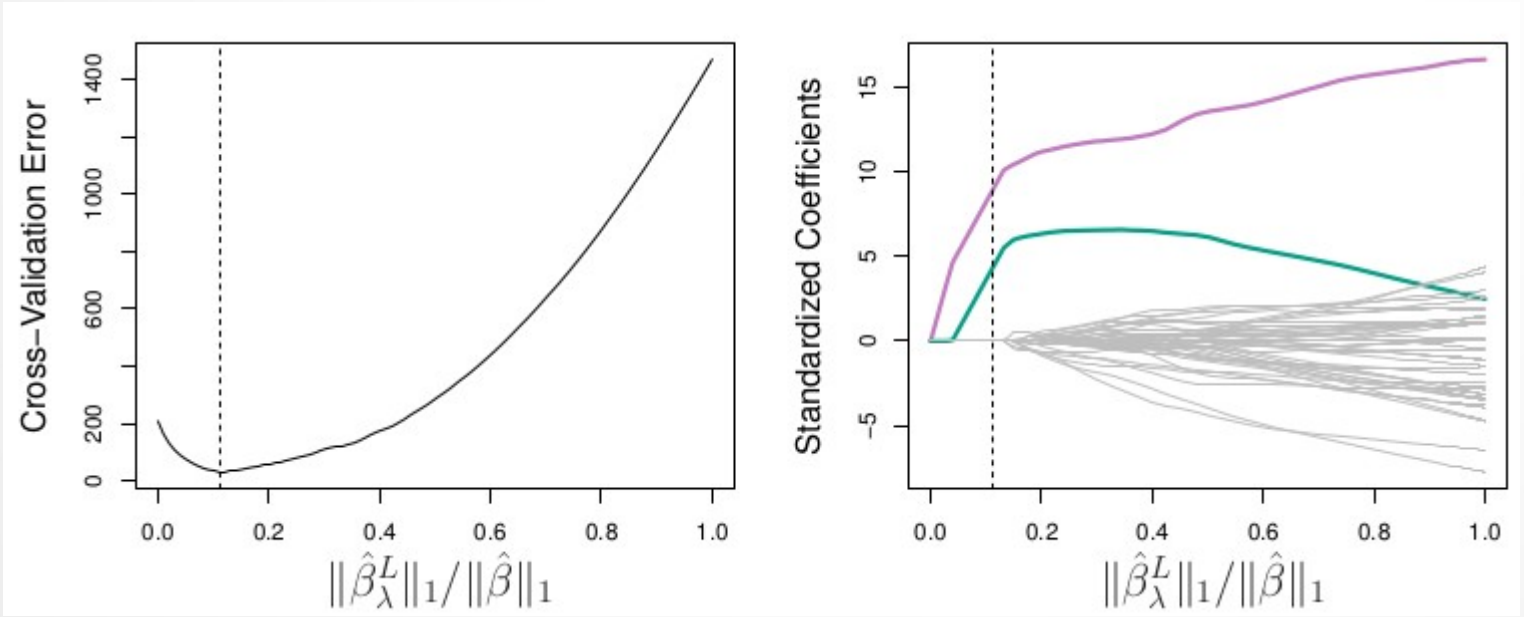
SOLUTION:

- k-fold cross-validation (e.g split full data set into 10 parts/folds, training:test = 1:9) 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 (full data set) and the selected value of the tuning parameter.

Selecting Tuning Parameter for Ridge Regression and Lasso



Ridge regression



The Lasso