STA 314: Statistical Methods for Machine Learning I

Lecture - Shrinkage regression

Xin Bing

Department of Statistical Sciences
University of Toronto

Logistics

- Tutorial after class
- Project document out by tonight
- Project group formation:
  - Sign-up form (team name, team members)
  - Quercus (team members)
  - Till Feb 1st, 11:59pm
  - Choose the data set
- Sign-up on kaggle starts on Feb 2, 10am, ends on Feb 6, 11:59pm.
  - only sign-up to the chosen data set
  - non-registered team name will be removed

Review

- Best subset selection
  - Great! But computationally unaffordable (choose from 2^p models)!
- Stepwise subset selection
  - Forward stepwise selection
  - Backward stepwise selection
  - Computationally affordable, but greedy approaches
- Are there better alternatives?
  - **Shrinkage methods**! In particular, the Lasso.

Shrinkage Methods / Regularization

- We can fit a model containing all p predictors using a technique that **constrains** or **regularizes** the coefficient estimates by shrinking the coefficient estimates towards zero.
- Shrinking the coefficient estimates can significantly reduce their variances.
- The two best-known techniques for shrinking the regression coefficients towards zero are the **ridge regression** and the **lasso**.

Ridge Regression

- Recall that the OLS fitting procedure estimates $\beta_0, ..., \beta_p$ using the values that minimize

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

- The **ridge regression** estimates $\beta_0, \ldots, \beta_p$ using the values that minimize

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

where $\lambda \geq 0$ is a tuning (regularization) parameter, to be determined later.

Comments

$$\hat{\boldsymbol{\beta}}_\lambda^R = \underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2.$$

- We usually denote the ridge regression estimator by $\hat{\boldsymbol{\beta}}_\lambda^R$, because different λ's produce distinct estimators.
- The term $\lambda\sum_{j=1}^{p}\beta_j^2$ is called a **shrinkage / regularization penalty**, which shrinks the estimates of each β_j towards 0.
- We usually do not penalize the intercept β_0.
- Comparing to the OLS estimator, the ridge regression finds the coefficient estimate of $\boldsymbol{\beta}$ that has small entries (toward 0) by affording a slightly larger RSS. The balance is controlled by λ.

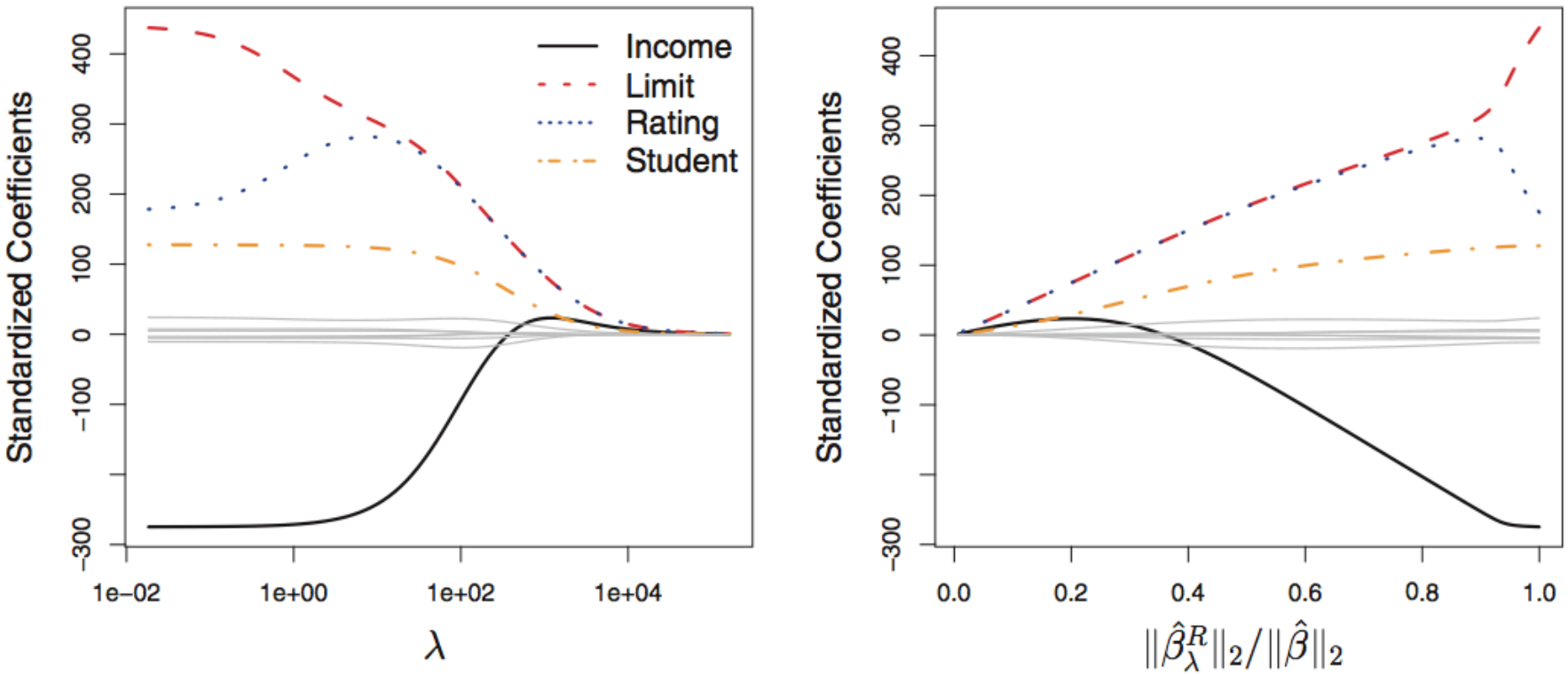
More Comments

- Selecting a good value for λ is critical. For $\lambda = 0$, the ridge estimator of $\boldsymbol{\beta}$ coincides with the OLS estimator. Cross-validation could be used to select λ.
- In practice, we recommend the standardized predictors for ridge regression, using the formula

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

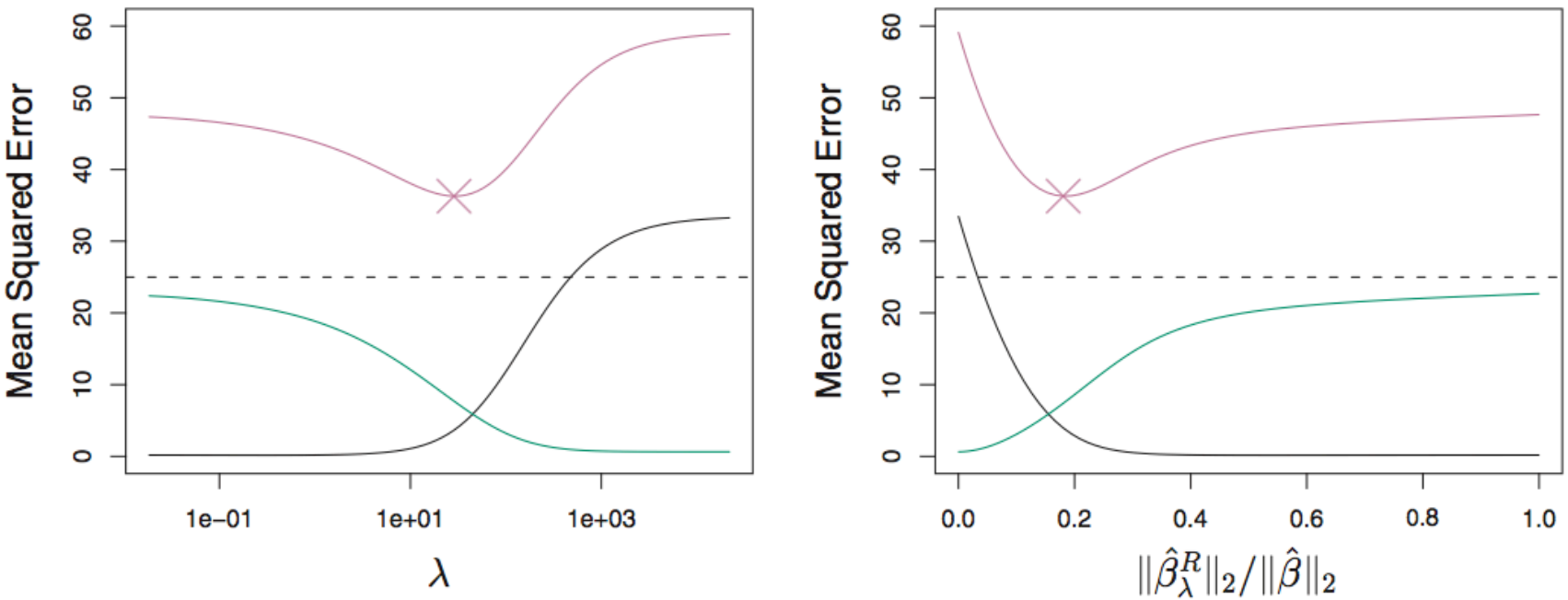
  All standardized predictors have standard deviation equal to one.

Credit Card Data Example



- In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the 10 variables, plotted as a function of λ.
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but we now display $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$, where $\hat{\beta}$ denotes OLS estimator.

Ridge Regression Improves Over OLS in terms of MSE



Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression. The dashed lines indicate the smallest possible MSE.

Advantages of Ridge Regression

- Ridge does a better job for prediction than the OLS approach by reducing the coefficient estimates.
  - Ridge reduces the variance of fitted model by trading off the bias

- Ridge regression is computationally efficient (for a given λ), comparable to the OLS approach.
  In particular, it is (much) faster than the best subset selection.

Limitation of Ridge Regression

- Can we use ridge regression for variable selection (excluding features that are not important by setting their estimates to 0)?

  No, it tends to include all p features in the fitted model!
  Meaning that the fitted model still has p non-zero coefficients.

The Lasso

- Different from ridge, lasso shrinks the coefficients by penalizing their absolute values.
- Specifically, the lasso coefficients, $\hat{\boldsymbol{\beta}}_\lambda^L$, minimize the quantity

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

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

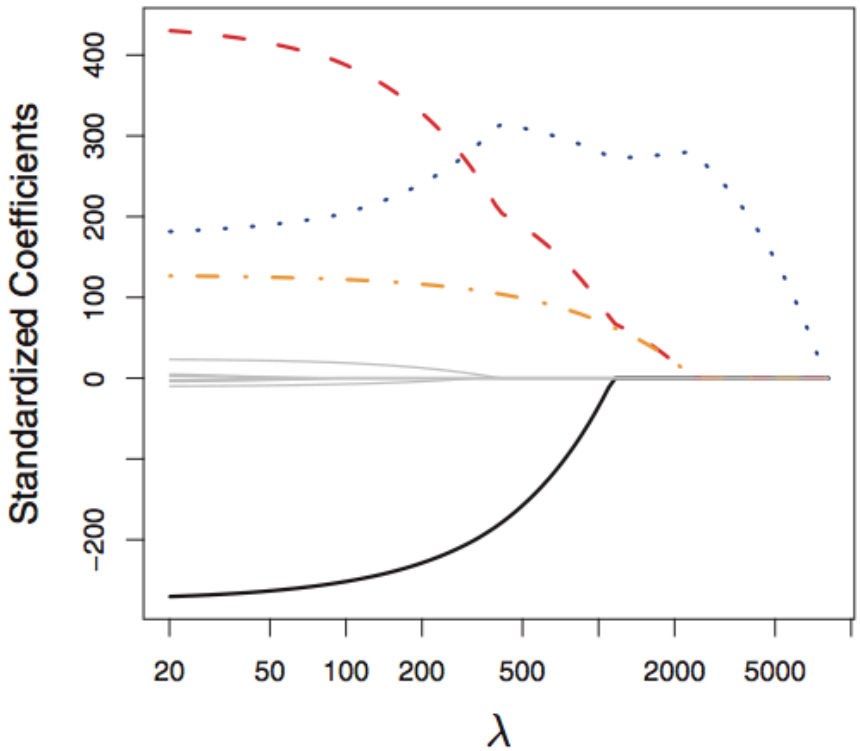
- Different from the ridge regression that uses the ℓ_2 penalty $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p}\beta_j^2$, lasso uses the ℓ_1 penalty

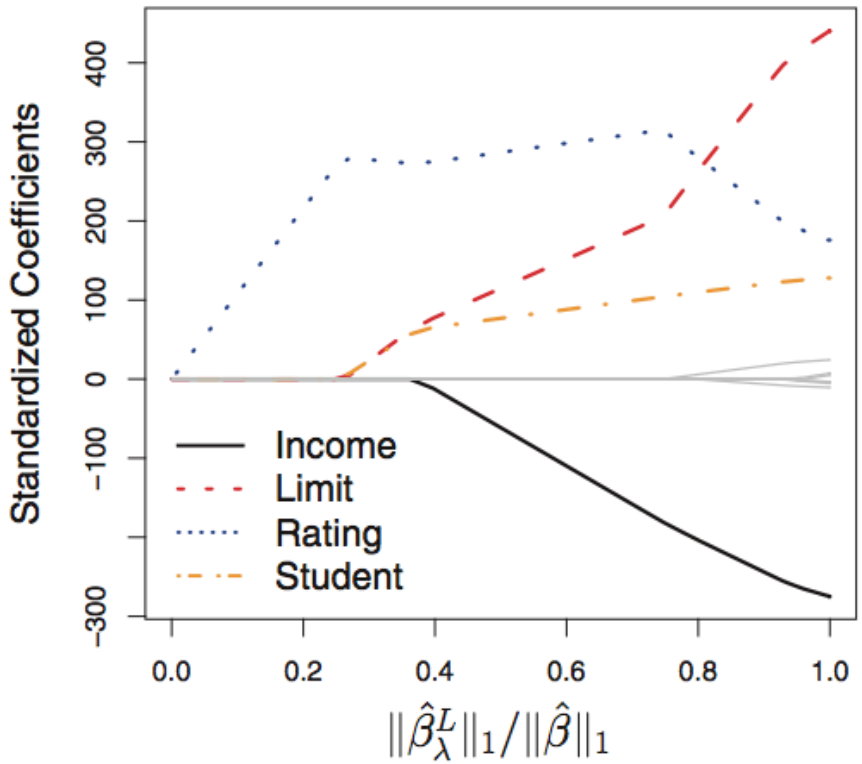
$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|.$$

More Comments

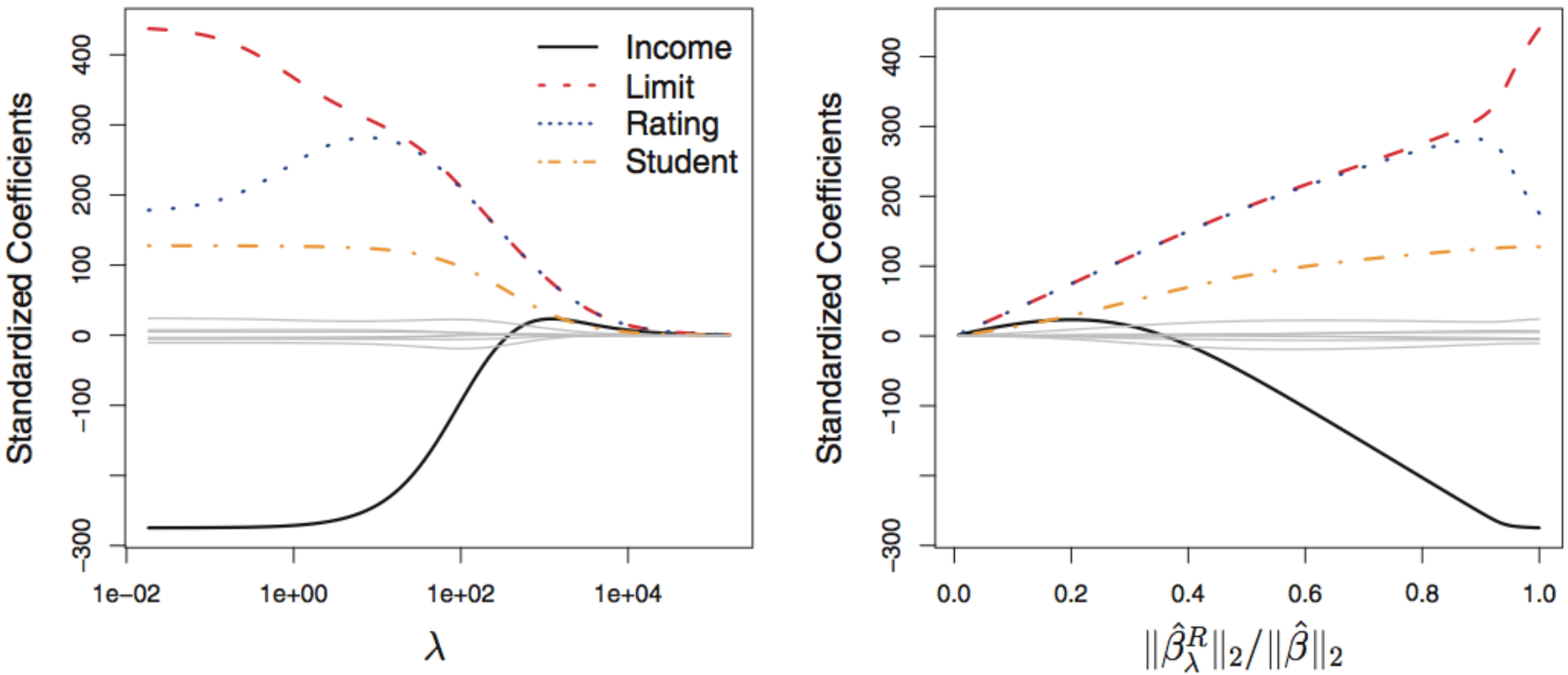
- Similar to ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be **exact zero** when the tuning parameter λ is sufficiently large.
- Therefore, the lasso performs variable selection.
- We say that the lasso yields a **sparse model** if the fitted model involves only a subset of the variables.
- Similar to ridge regression, selecting a good value of the regularization parameter λ for the lasso is critical; cross-validation is again the method of choice.

Credit Card Data Example





Credit Card Data Example



- In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the 10 variables, plotted as a function of λ.
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but we now display $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$, where $\hat{\beta}$ denotes OLS estimator.

Magic of the Lasso

Why does the lasso, unlike ridge regression, yield coefficient estimates that have exact zero?

Another Formulation for Ridge Regression and Lasso

The lasso and ridge regression coefficient estimates solve the problems

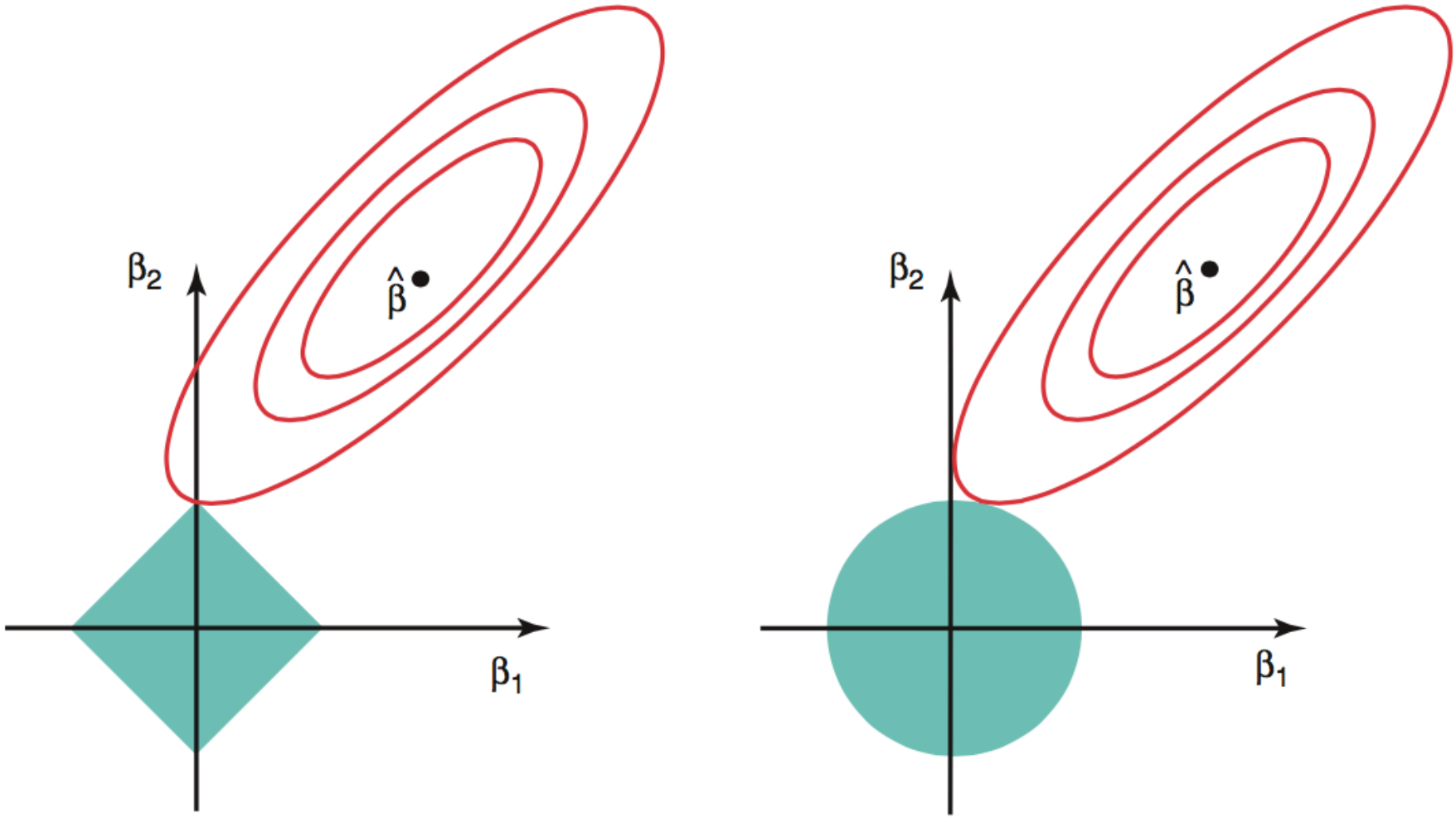
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

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

Here $s \geq 0$ is some regularization parameter (related with λ before).

Understand why the Lasso yields zero estimates

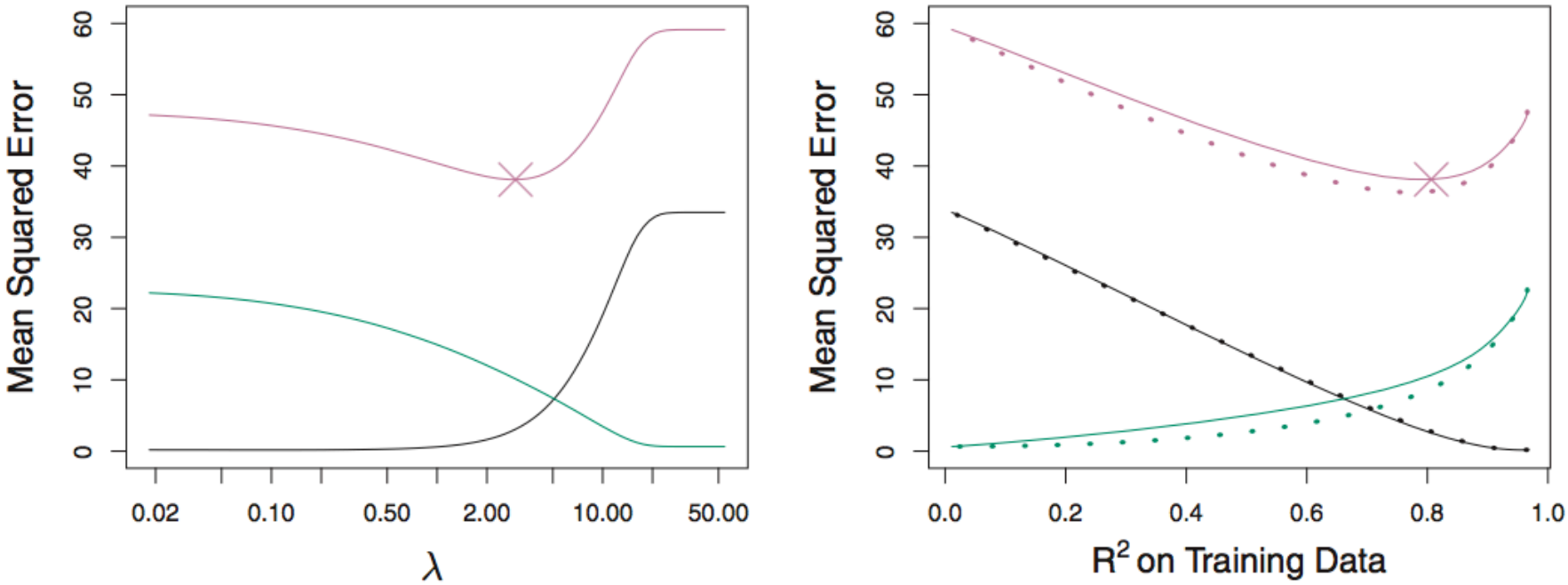


The solid areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.

Lasso vs Ridge

- The ability of yielding a **sparse** model is a big advantage of Lasso comparing to Ridge.
- A more sparse model means more interpretability!
- What about their prediction performance?

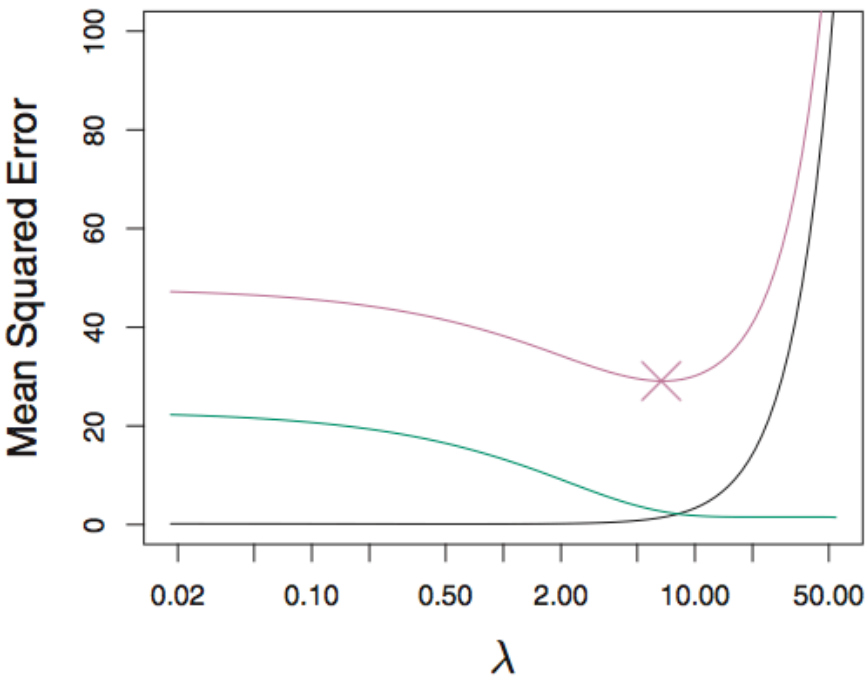
Comparing the MSE of Lasso and Ridge

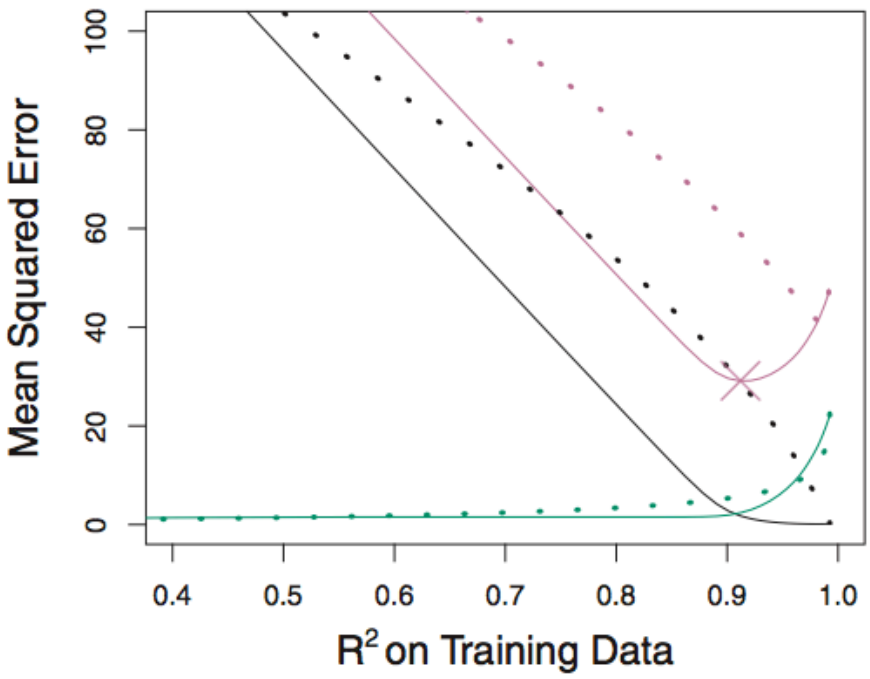


Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.
Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their R^2 on the training data, as a common form of indexing.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

- *When the true coefficients are non-sparse, ridge and lasso have the same bias but ridge has a smaller variance hence a smaller MSE.*

Another Case





- *When the true coefficients are sparse, Lasso outperforms ridge regression of having both a smaller bias and a smaller variance.*

Conclusions on Lasso relative to Ridge

- Neither ridge regression nor the lasso universally dominates the other.
- In general, the lasso performs better when the response is only related with a relatively small number of predictors.
- As the ridge regression, when the OLS estimates have high variance, the lasso solution has smaller variance at the expense of a small increase in bias, and consequently can lead to more accurate predictions.
- Unlike ridge regression, the lasso performs variable selection, and hence yields models that are easier to interpret.

A simple example of the shrinkage effects of ridge and lasso

- Assume that $n = p$ and $\mathbf{X} = \mathbf{I}_n$. We force the intercept term $\beta_0 = 0$.
- In this way,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{bmatrix}.$$

- We assume

$$\mathbb{E}[\epsilon_j] = 0, \qquad \mathbb{E}[\epsilon_j^2] = \sigma^2, \qquad \forall j \in [p] := \{1, 2, \ldots, p\}$$

The OLS estimator

- The OLS approach is to find $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p} (y_j - \beta_j)^2.$$

This gives the OLS estimator

$$\hat{\beta}_j = y_j, \qquad \forall j \in [p].$$

The ridge estimator

- The ridge regression is to find $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

This leads to the ridge estimator

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda}, \qquad \forall j \in [p].$$

Since $\lambda \geq 0$, the magnitude of each estimated coefficient is shrinked toward 0.

The lasso estimator

- The lasso is to find $\beta_1, \ldots, \beta_p$ that minimize

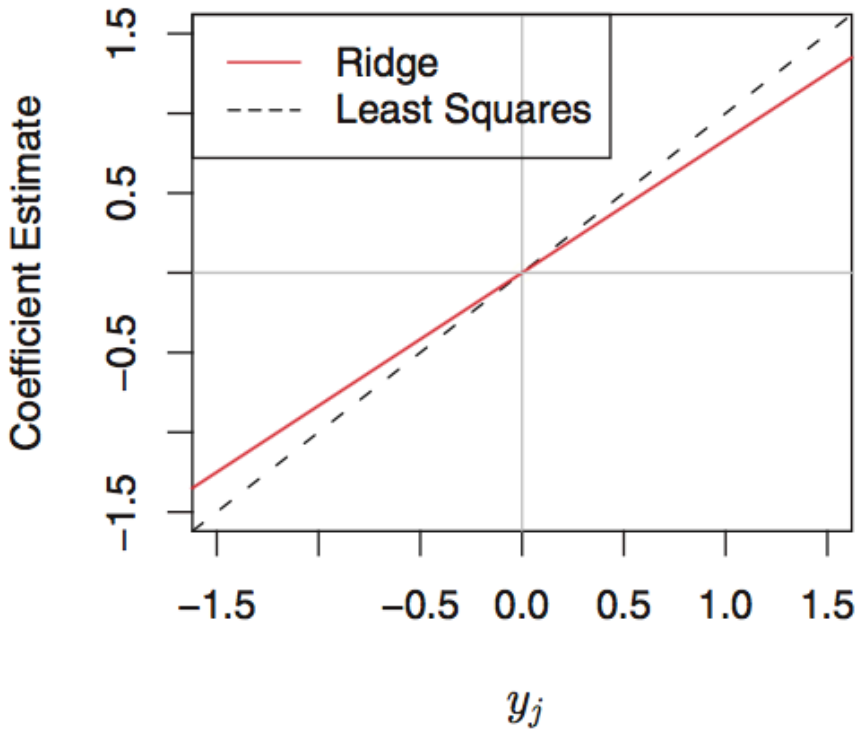
$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

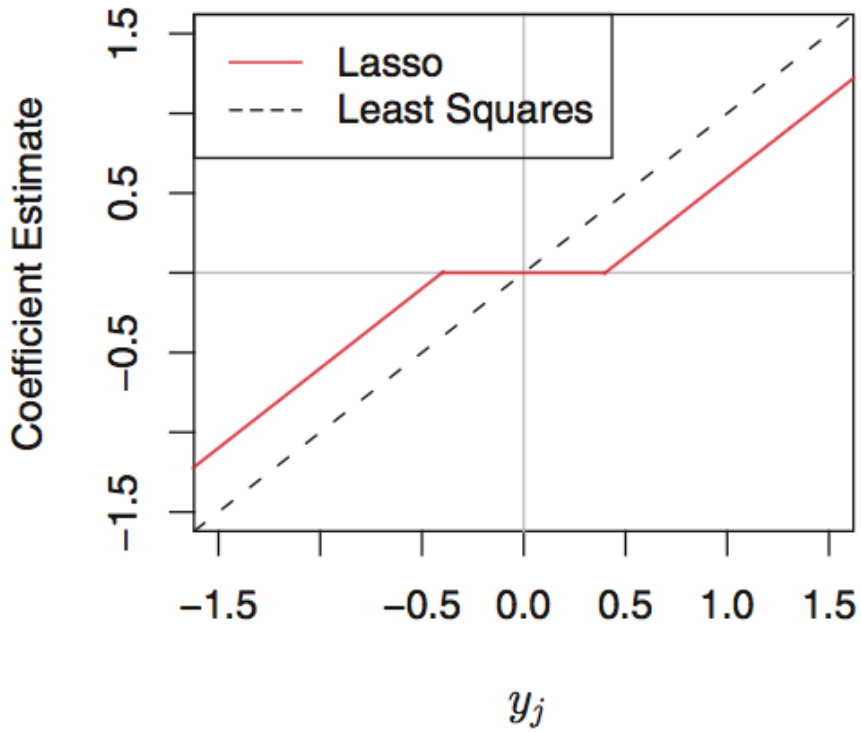
This gives estimator

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

The estimated coefficients from Lasso are also shrinked. The above shrinkage is known as the **soft-thresholding**.

An illustrative figure ($\lambda = 1$)





Bias and Variance of the OLS

Recall

$$y_j = \beta_j + \epsilon_j, \qquad \forall j \in [p].$$

For any $j \in [p]$, the OLS estimator $\hat{\beta}_j = y_j$ satisfies

- **Bias**:

$$\mathbb{E}[\hat{\beta}_j] = \mathbb{E}[y_j] = \mathbb{E}[\beta_j + \epsilon_j] = \beta_j$$

- **Variance**:

$$\mathrm{Var}(\hat{\beta}_j) = \mathrm{Var}(\epsilon_j) = \sigma^2$$

MSE of the OLS

- **Mean squared error** of the jth coefficient:

$$\mathbb{E}\left[\left(\hat{\beta}_j - \beta_j\right)^2\right] = \left(\mathbb{E}[\hat{\beta}_j] - \beta_j\right)^2 + \mathrm{Var}(\hat{\beta}_j) = \sigma^2$$

- **Mean squared error** of all p coefficients:

$$\mathbb{E}\left[\sum_{j=1}^{p}\left(\hat{\beta}_j - \beta_j\right)^2\right] = p\sigma^2.$$

Bias and Variance of the Ridge

Recall

$$y_j = \beta_j + \epsilon_j, \qquad \forall j \in [p].$$

For any $j \in [p]$, the ridge estimator with tuning parameter λ,

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda},$$

satisfies

- **Bias**:

$$\mathbb{E}[\hat{\beta}_j^R] = \mathbb{E}\left[\frac{y_j}{1+\lambda}\right] = \mathbb{E}\left[\frac{\beta_j + \epsilon_j}{1+\lambda}\right] = \frac{\beta_j}{1+\lambda}.$$

- **Variance**:

$$\mathrm{Var}(\hat{\beta}_j^R) = \mathrm{Var}\left(\frac{\epsilon_j}{1+\lambda}\right) = \frac{\sigma^2}{(1+\lambda)^2}$$

MSE of the ridge

- **Mean squared error** of the jth coefficient:

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\beta}_j^R - \beta_j\right)^2\right] &= \left(\mathbb{E}[\hat{\beta}_j^R] - \beta_j\right)^2 + \mathrm{Var}(\hat{\beta}_j^R) \\ &= \left(\frac{\beta_j}{1+\lambda} - \beta_j\right)^2 + \frac{\sigma^2}{(1+\lambda)^2} \\ &= \frac{\lambda^2\beta_j^2}{(1+\lambda)^2} + \frac{\sigma^2}{(1+\lambda)^2}.\end{aligned}$$

  Recall that $\mathbb{E}[(\hat{\beta}_j - \beta_j)^2] = \sigma^2$.

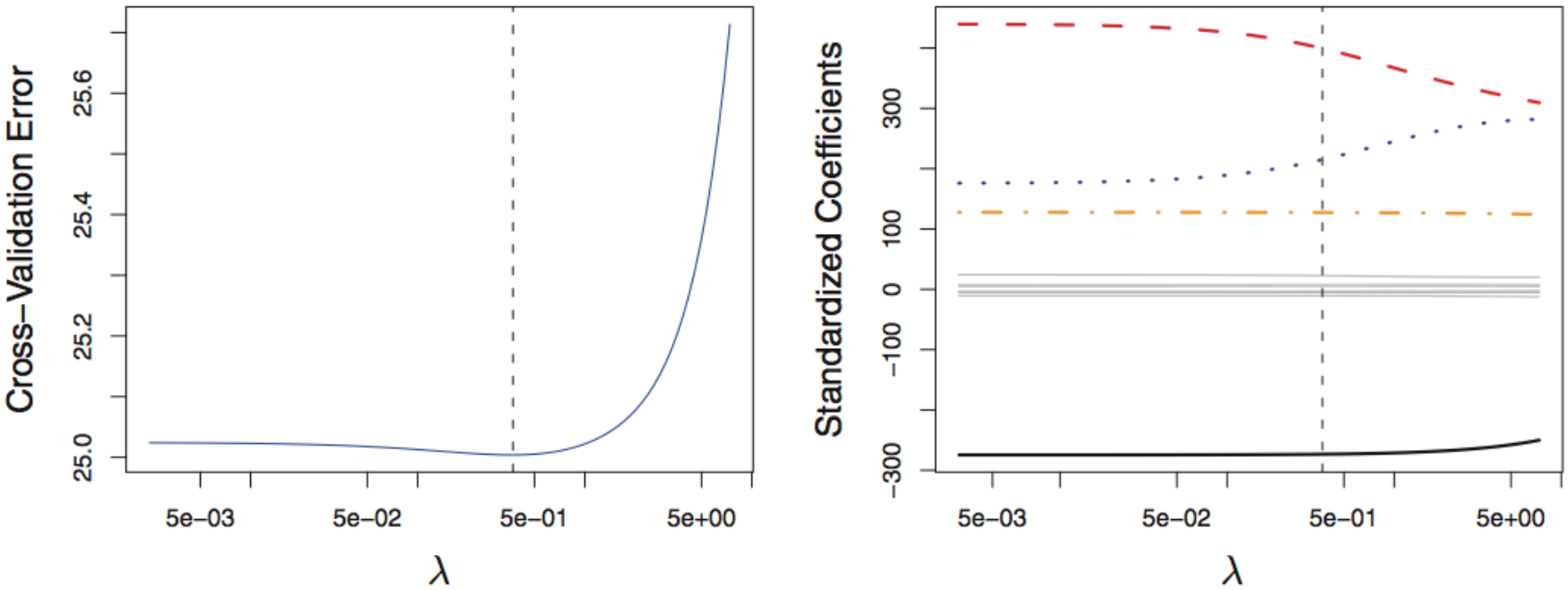
- **Mean squared error** of all p coefficients:

$$\mathbb{E}\left[\sum_{j=1}^{p}\left(\hat{\beta}_j^R - \beta_j\right)^2\right] = \frac{\lambda^2}{(1+\lambda)^2}\sum_{j=1}^{p}\beta_j^2 + \frac{p\sigma^2}{(1+\lambda)^2}.$$

On selecting the tuning parameter

- Similar as the subset selection, for ridge and lasso, we require a systematic way of choosing the best model under a sequence of fitted models (from different choices of λ)
  - Equivalently, we require a method to select the optimal value of the tuning parameter λ.
- Cross-validation: we choose a grid of λ, and compute the cross-validation error rate for each value of λ.
- We then select the λ_* for which the cross-validation error is smallest.
- Finally, the model is re-fitted by using all observations and λ_*.

Credit Card Data Example



Cross-validation errors that result from applying ridge regression to the Credit data set for various choices of λ.

More choices of penalties

- There are many other penalties in addition to the ℓ_2 and ℓ_1 norms used by ridge and lasso.
  - the elastic net:

$$\min_{\boldsymbol{\beta}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left[(1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2\right]$$

    for some tuning parameters $\lambda \geq 0$ and $\alpha \in [0, 1]$.

    - The ridge corresponds to $\alpha = 1$
    - The Lasso corresponds to $\alpha = 0$.

The group lasso

- If we suspect the model is nonlinear in X_1 or X_2, we can add quadratic terms, say

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \epsilon.$$

The **group lasso** estimator minimizes

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left(\sqrt{\beta_1^2 + \beta_2^2} + \sqrt{\beta_3^2 + \beta_4^2} \right).$$

In this penalty, we view β_1 and β_2 (coefficient of X_1 and X_1^2) as if they belong to the same group. The group Lasso can shrink the parameters in the same group (both β_1 and β_2) exactly to 0 simultaneously.

- There are a lot more penalties out there .....

Regularization in more general settings

- The ridge and lasso regressions are not restricted to the linear models.

- The idea of penalization is generally applicable to almost all parametric models.

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; L(\boldsymbol{\beta}, \mathcal{D}^{train}) + \lambda \cdot Pen(\boldsymbol{\beta}).$$

  - OLS: $L(\boldsymbol{\beta}, \mathcal{D}^{train}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, $Pen(\boldsymbol{\beta}) = 0$.
  - Ridge: $L(\boldsymbol{\beta}, \mathcal{D}^{train}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, $Pen(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$.
  - Lasso: $L(\boldsymbol{\beta}, \mathcal{D}^{train}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, $Pen(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$.

  - In principal,
    - L can be any loss function, i.e. negative likelihood, 0-1 loss.
    - Pen could be any penalty function.