STA 314: Statistical Methods for Machine Learning I

Lecture 4 - Model selection under linear models: subset selection and shrinkage regression

Xin Bing

Department of Statistical Sciences
University of Toronto

Review

We have learned two approaches for model selection when we don't have $\mathcal{D}_{test}$:

- Avoid estimating the expected MSE by adjusting the training error to account for the model complexity:
  - Mallow's C_p
  - AIC
  - BIC
  - adjusted R^2
- Directly estimate the expected MSE via data-splitting:
  - validation set approach
  - cross-validation approach

Application to model selection in linear models

Recall that we have the following alternatives to the OLS using all predictors:

- **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 the OLS approach on the identified set of predictors.
  - **Best Subset Selection**
  - **Stepwise Selection**
- **Shrinkage Regression**
  - **Ridge**
  - **Lasso**
- **Dimension Reduction**. Later after PCA.

Best Subset Selection

Example

Suppose we have access to i.i.d. samples of the response Y and the features

$$X = (X_1, X_2, X_3).$$

For fitting a regression that is linear in X, what are the all possible subsets?

Best Subset Selection

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

- Step 2 identifies the best model for each subset size. RSS can be used here. Why?
- In Step 3, can we use *RSS* or R^2?

Best Subset Selection

- The same approach can be used for other types of models, such as logistic regression (RSS replaced by deviance).
- However! For best subset selection, we need to fit and compare

$$\binom{p}{0} + \binom{p}{1} + \binom{p}{2} + \cdots + \binom{p}{p} = 2^p$$

models.

Forward Stepwise Selection

Example (Revisited)

Suppose we have access to i.i.d. samples of the response Y and the features

$$X = (X_1, X_2, X_3).$$

What are the models we consider for forward stepwise?

Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ 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 $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward Stepwise Selection

- **Pros:** It has computational advantage over best subset selection. In the kth iteration, we fit and compare $(p-k)$ models. In total,

$$1 + \sum_{k=0}^{p-1} (p-k) = 1 + \frac{p(p+1)}{2}$$

models are considered, *much fewer* than 2^p models.

- **Cons:** It is a greedy procedure!
So not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

The Credit Card Data

# 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

Example (Revisited)

Suppose we have access to i.i.d. samples of the response Y and the features

$$X = (X_1, X_2, X_3).$$

What are the models we consider for backward stepwise?

Backward Stepwise Selection

Algorithm 6.3 *Backward stepwise selection*

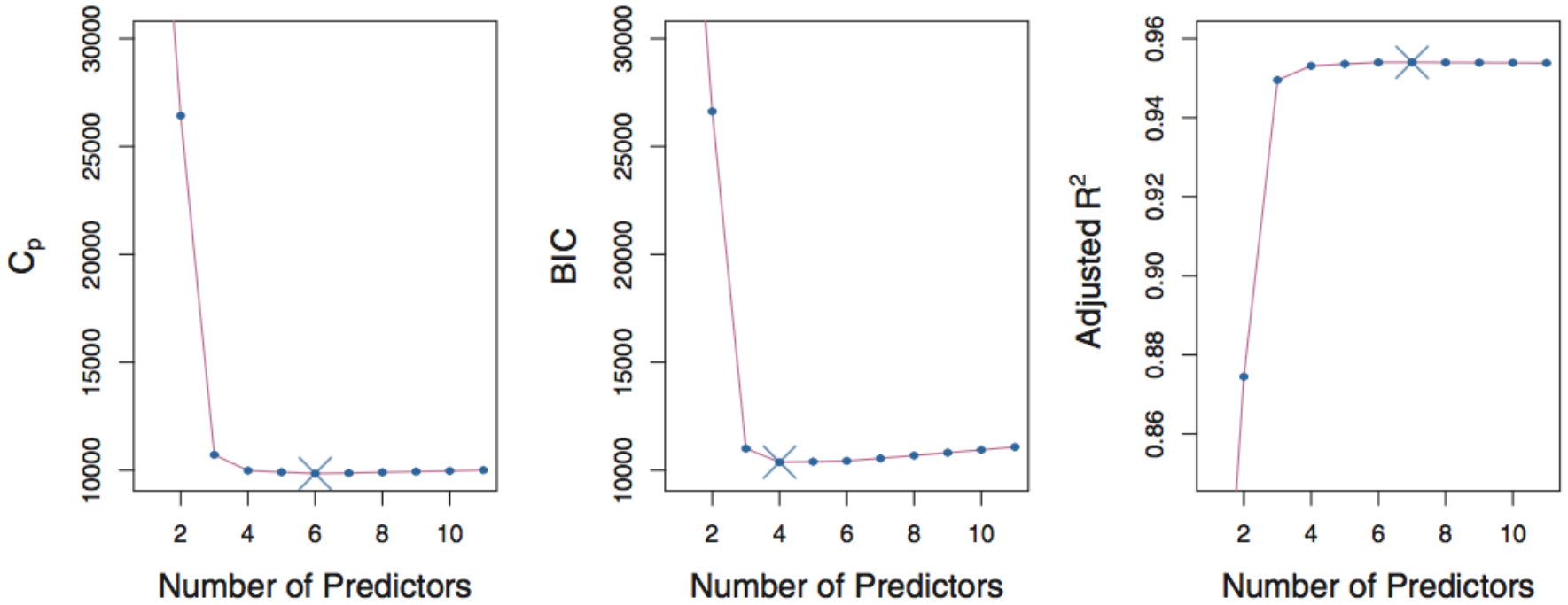
1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Backward Stepwise Selection

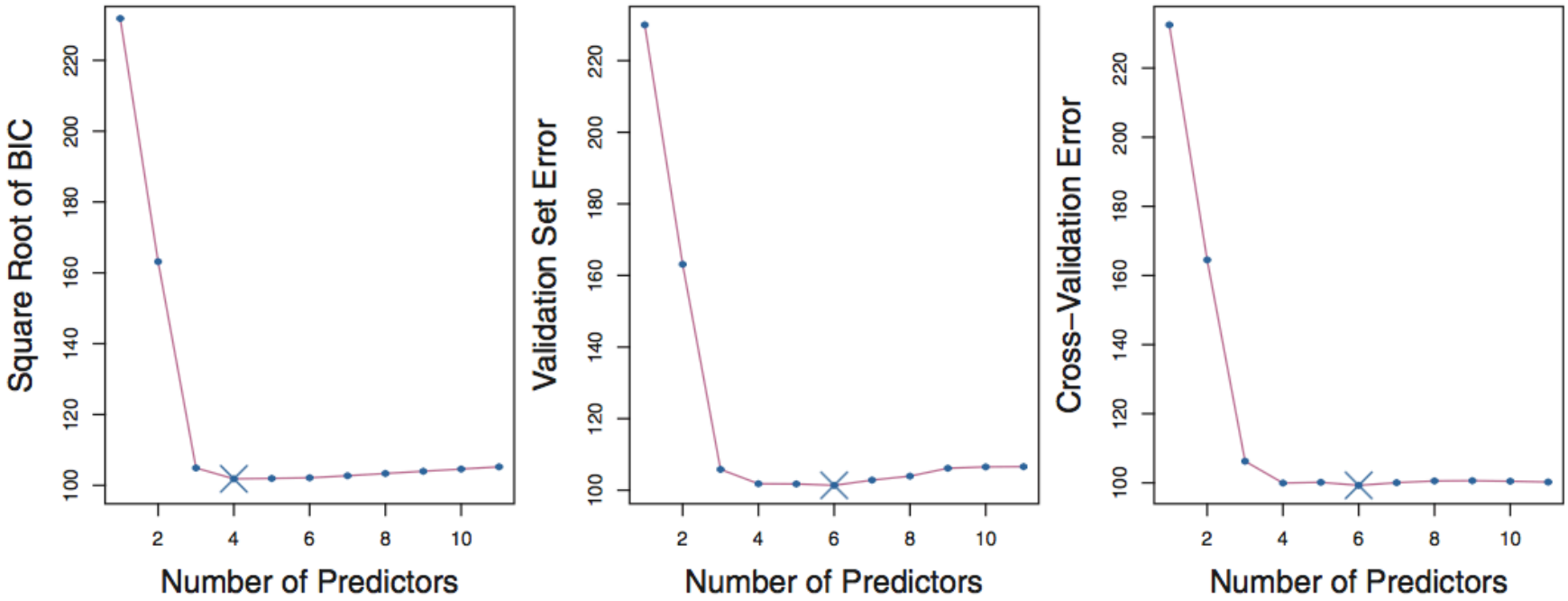
- For backward stepwise selection, we also compare $1 + p(p+1)/2$ models, much fewer than 2^p models.

- Still a greedy approach!
  It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

The Credit Card Data: best subset selection via Mallow's C_p, BIC and adjusted R^2



The Credit Card Data: model selection via sample-splitting



Summary on subset selection

- Best subset selection will select the best model, as long as computation is affordable.
- Forward / Backward stepwise selection is computationally fast, but is not guaranteed to find the best model.
- What should we do in practice?

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

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

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

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

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

Comments

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

- 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.

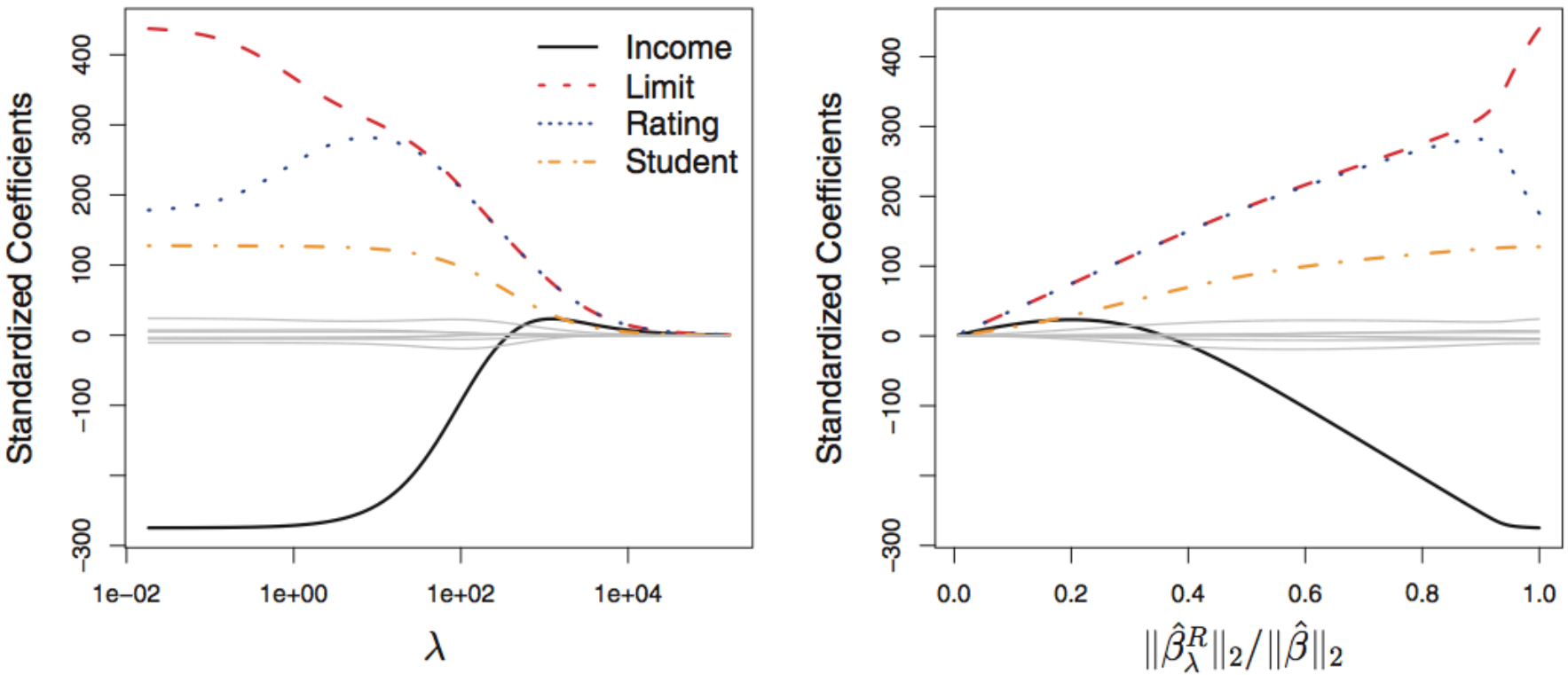
More Comments

- 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

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

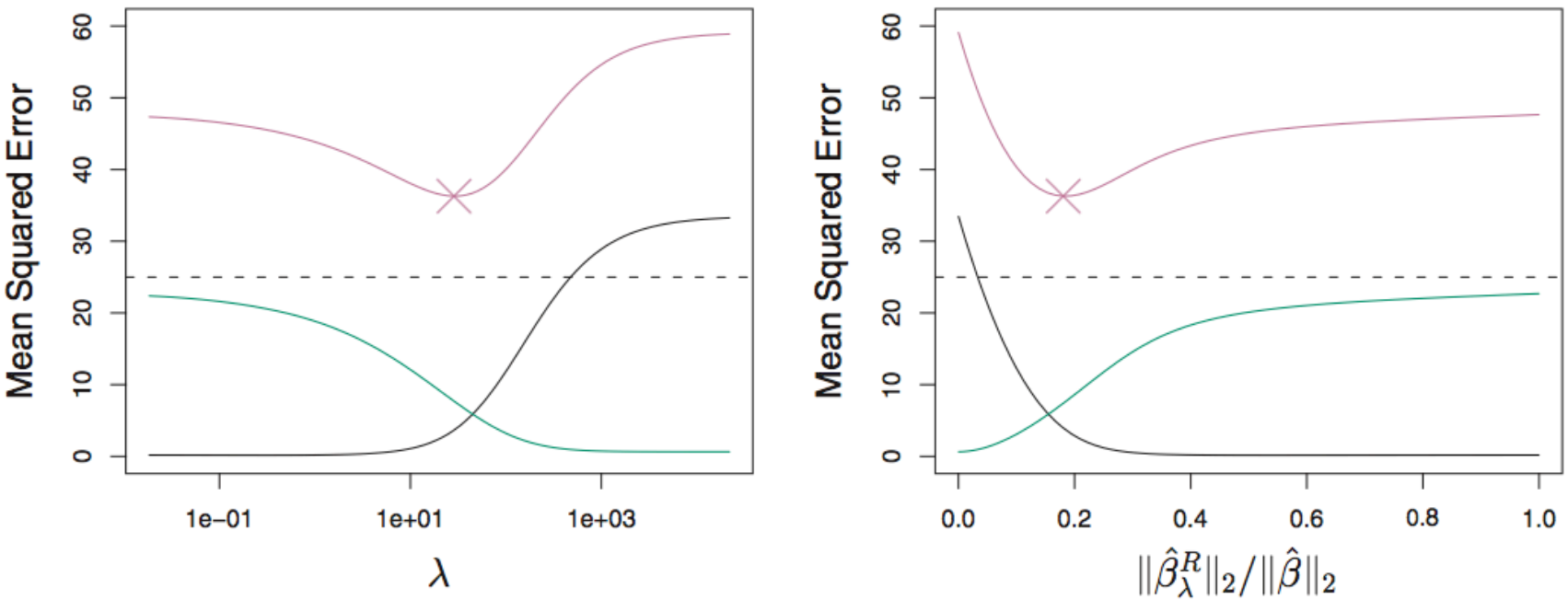
  All standardized predictors have standard deviation equal to one.

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 has substantial computational advantages over 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!

  So, the resulting fitted model is difficult to interpret.

The Lasso

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

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

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

- Different from the ridge regression that uses the ℓ_2 penalty

$$\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p}\beta_j^2,$$

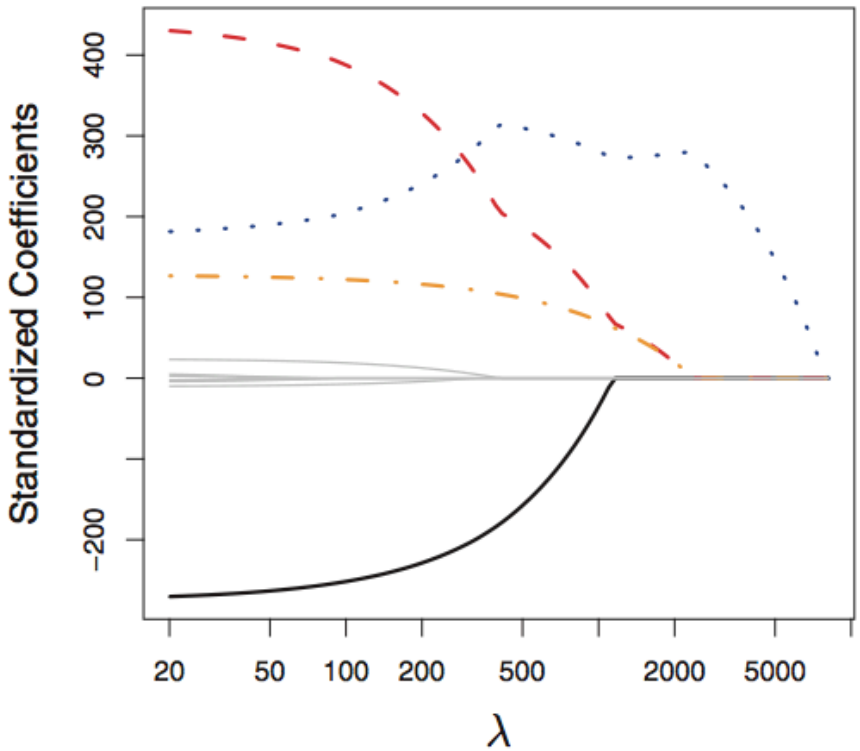
  lasso uses the ℓ_1 penalty

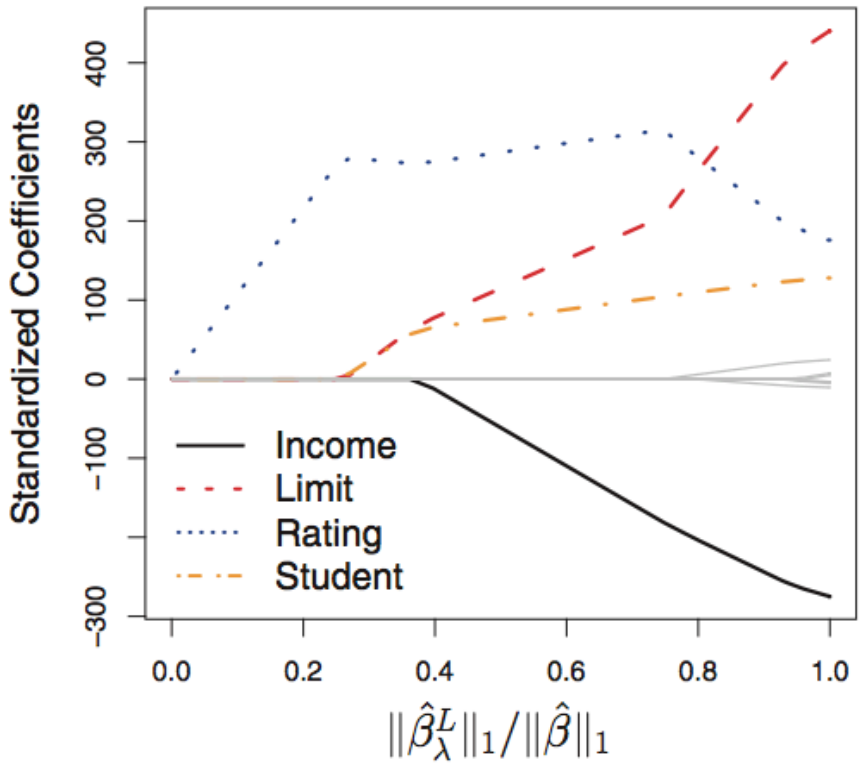
$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|.$$

More Comments

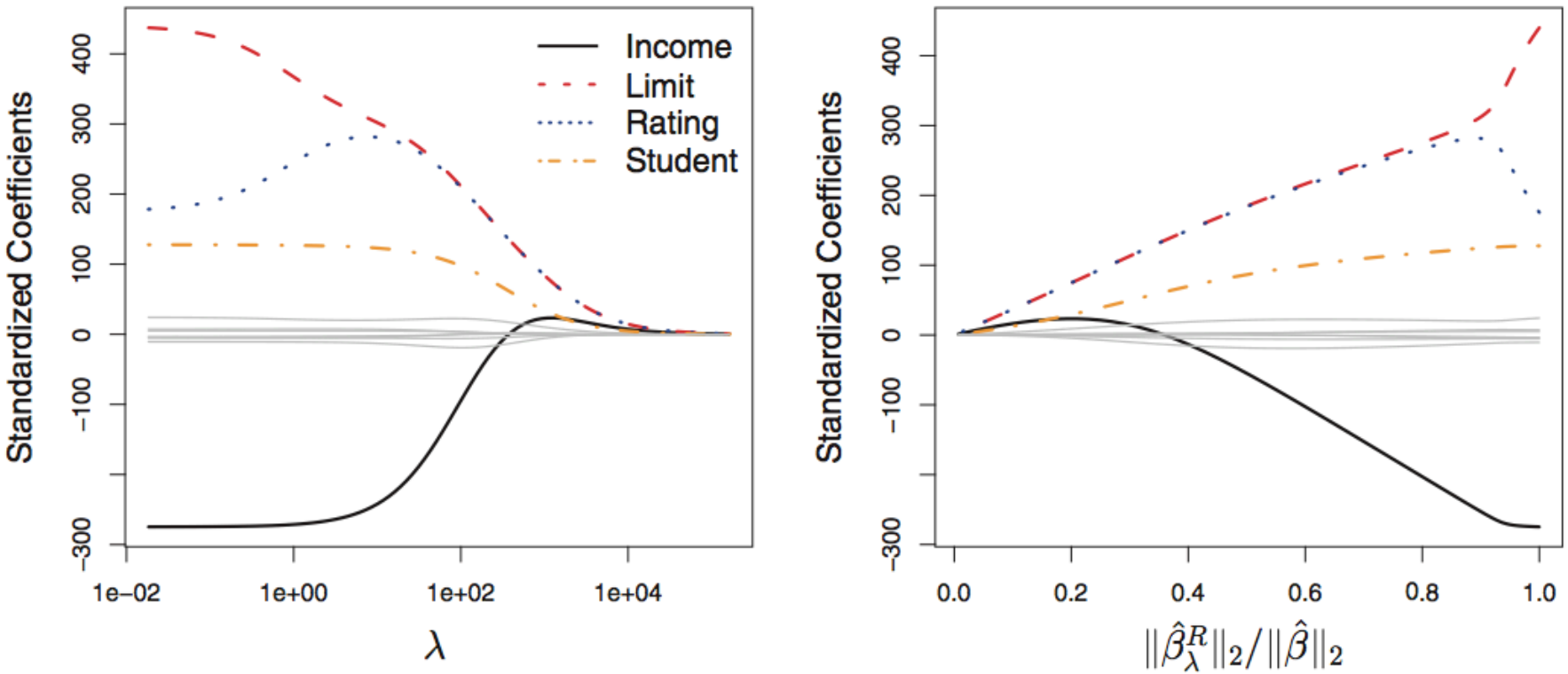
- Similar to ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be **exact zero** when the tuning parameter λ is sufficiently large.
- Therefore, the lasso performs 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.