Lecture 14: Linear Model Selection and Regularization

Nayel Bettache

Department of Statistical Science, Cornell University

Shrinkage Methods

- 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.
- Shrinking the coefficient estimates can significantly reduce their variance.
- The two best-known techniques for shrinking the regression coefficients towards zero are **ridge regression** and the **lasso**.

Ridge Regression

- The least squares 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 $\beta_0, ..., \beta_p$ 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 parameter, to be determined later.
- The term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called a shrink penalty, which shrinks the estimates of β_i towards 0.
- Ridge regression seeks coefficient estimates that making *RSS* small, and meanwhile shrinks β_j towards 0. The relative importance of two terms is controlled by λ.

More Comments

- We usually denote the ridge regression estimator by $\hat{\beta}_{\lambda}^{R}$, because different λ produces different estimators.

- Selecting a good value for λ is critical. Later, we use cross-validation to select λ. Some other criterions discussed in the previous section, such as AIC, BIC can be also used.

- We usually do not penalize β_0.

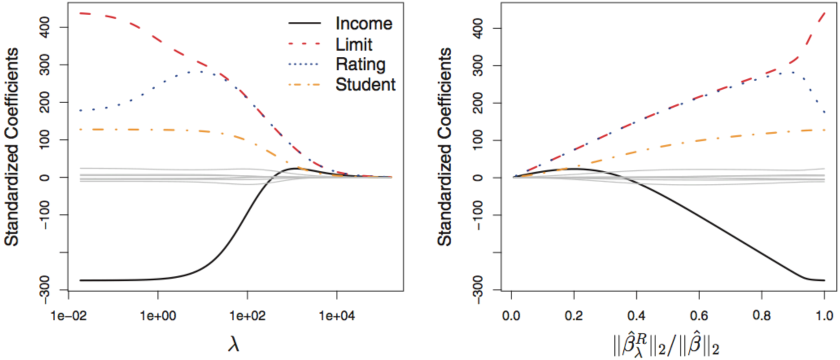
- The ridge regression coefficient estimates are not equivariant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.

- In practice, we recommend the standardized predictors, using the formula

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

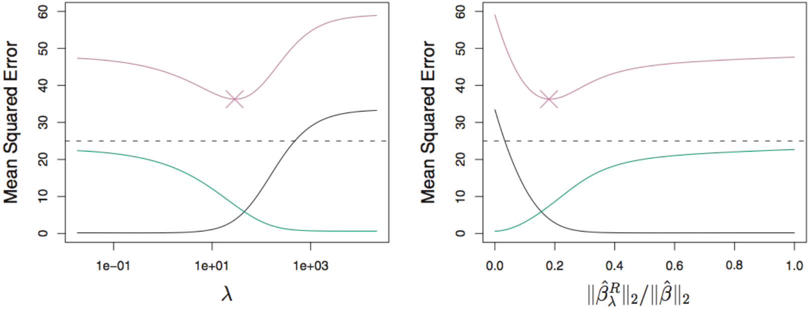
  All of the standardized predictors will have a standard deviation of one.

Credit Card Data Example



- 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 we now display $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$, where $\hat{\beta}$ denotes LSE.

Ridge Regression Improves Over Least Squares



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

Ridge regression also has substantial computational advantages over best subset selection. How to apply ridge regression?

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. So, the model interpretation is more difficult.

- The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

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

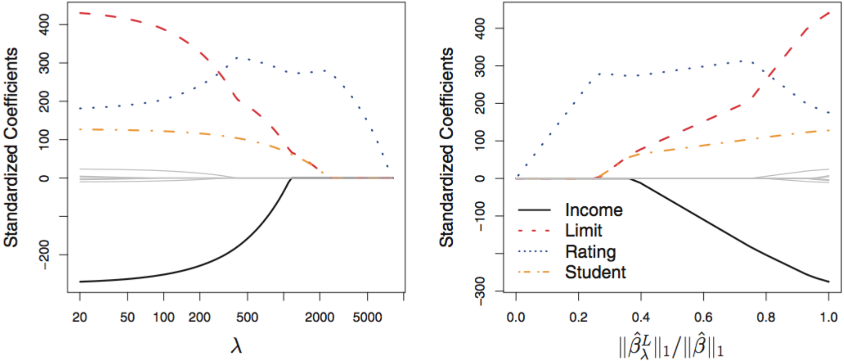
  where $\lambda \geq 0$ is a tuning parameter, to be determined later.

- Different from the ridge regression which uses L_2 penalty $\|\beta\|_2^2$, lasso uses L_1 penalty $\|\beta\|_1$.

More Comments

- As with 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.

Credit Card Data Example



Another Formulation for Ridge Regression and Lasso

Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?

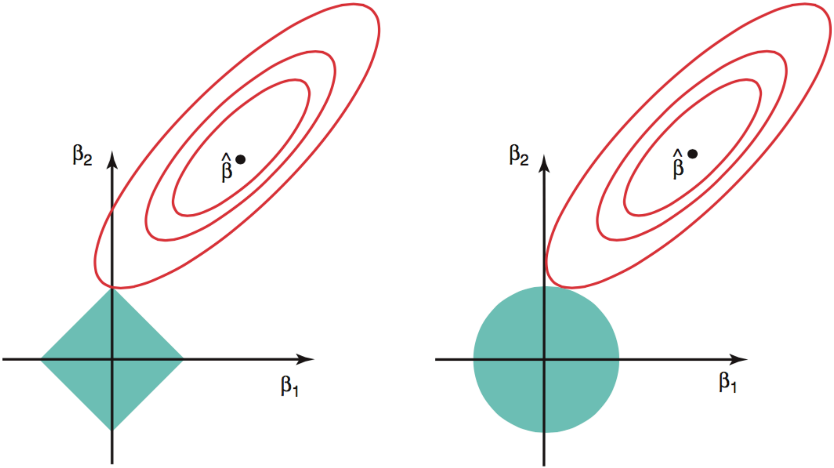
The lasso and ridge regression coefficient estimates solve the problems

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

and

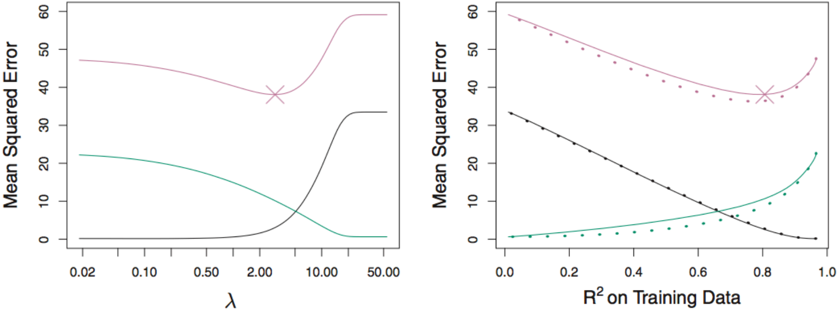
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

The Variable Selection Property of the Lasso



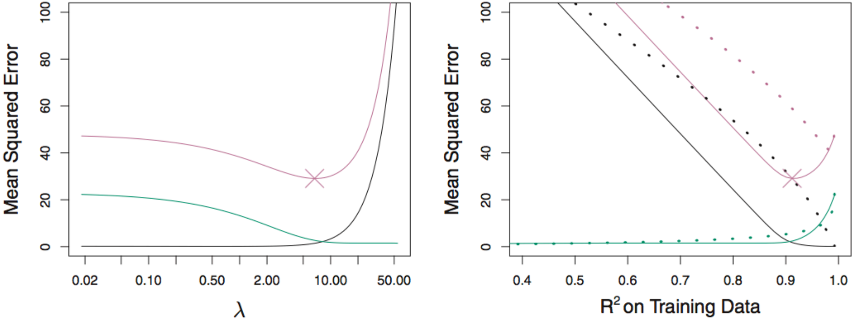
The solid areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.

Comparing the Lasso and Ridge Regression



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 R2 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.

Another Case



Lasso outperforms ridge regression when the true model is sparse.

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.
- As with ridge regression, when the least squares estimates have excessively high variance, the lasso solution can yield a reduction in variance at the expense of a small increase in bias, and consequently can generate more accurate predictions.
- Unlike ridge regression, the lasso performs variable selection, and hence results in models that are easier to interpret.
- There are very efficient algorithms for fitting both ridge and lasso models; in both cases the entire coefficient paths can be computed with about the same amount of work as a single least squares fit.