DSO530 Statistical Learning Methods

Lecture 6 part II : Shrinkage Methods

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
xint@marshall.usc.edu

Definition:
Bias: error introduced by approximating real life problem w/ much simpler model.
Variance: Amount by which estimated model would change if we used a different training set.

Bias-variance trade-off (optional)

usually $\varepsilon \sim N(0, \sigma^2)$

- Suppose the true relation between x and y is

$$y = f(x) + \varepsilon$$

- Based on training set $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, we construct $\hat{f}$ to estimate f
- for a new pair of observation (x_0, y_0), we predict y_0 using $\hat{f}(x_0)$, and the discrepancy is $y_0 - \hat{f}(x_0)$
- Then the expected test MSE at x_0 is

What is the source of uncertainty?

Variance of error term.

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\varepsilon)$$

for fixed amt of data, as model complexity increases variance ↓ bias ↑ affected.

- where $Var(\hat{f}(x_0)) = E[\hat{f}(x_0) - E(\hat{f}(x_0))]^2$
- and $Bias(\hat{f}(x_0)) = E[\hat{f}(x_0) - f(x_0)]$

- The overall expected test MSE is computed by averaging $E\left(y_0 - \hat{f}(x_0)\right)^2$ over all possible values of x_0 in the test set.

as bias ↓, variance ↑. & vice versa

Shrinkage methods

↳ p>n, p≃n & not n>>p.

- 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 (with some cost in bias).
- Best known shrinkage methods: ridge regression and the lasso.
- The ridge regression coefficient estimates $\hat{\beta}_\lambda^R$ are the values that minimize

minimizes this function. Penalizes larger bj.

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

when λ=0, we get RSS formula. if λ→∞, bj→0. ⇓ null model.

where $\lambda \geq 0$ is a tuning parameter, and x_{ij} is the jth coordinate of the ith observation x_i.

- Lasso: find $\hat{\beta}_\lambda^L$ that minimizes

β₀ is never penalized.

only thing different from ridge.

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

In lasso, b_j is penalized, in ridge b_j is penalized.
[because predictors are treated equally in penalty term]

- In contrast to the usual least squares approach, rescaling the predictors is more important in shrinkage methods. Why?
- λ is an important penalty parameter that controls the amount of shrinkage.
- What happens when $\lambda = 0$ and $\lambda \to \infty$?
- How do we choose the tuning parameter λ ? Cross-validation
- Lasso tends to give sparser models compared to ridge (better for model interpretability), and it tends to perform better when the true model is sparse.
- But we do not know *a priori* which is better for prediction accuracy.

- In contrast to the usual least squares approach, rescaling the predictors is more important in shrinkage methods. Why?
- λ is an important penalty parameter that controls the amount of shrinkage.
- What happens when $\lambda = 0$ and $\lambda \to \infty$?
- How do we choose the tuning parameter λ ? Cross-validation
- Lasso tends to give sparser models compared to ridge (better for model interpretability), and it tends to perform better when the true model is sparse.
- But we do not know *a priori* which is better for prediction accuracy.

because only few predictors are used.

sparse model, many b_j's are 0.

Ridge regression does NOT give you sparse models

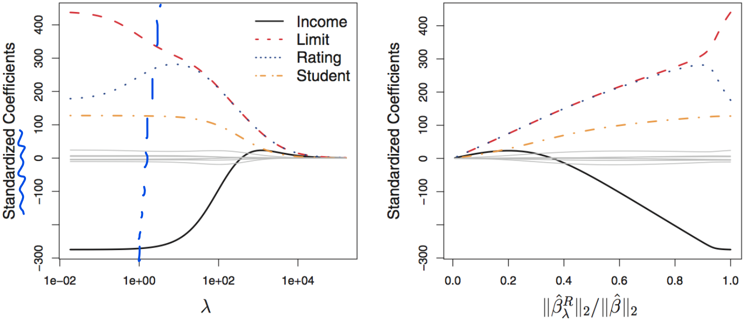


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* **Credit** *data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.*

- Note $\|\beta\|_2 = \sqrt{\beta_1^2 + \cdots + \beta_p^2}$
- $\hat{\beta}$ denotes the vector of least squares estimate

Bias-variance trade-off for ridge regression

If $\lambda \uparrow$ coeff $\downarrow$, $\therefore$ variability is less therefore variance decreases

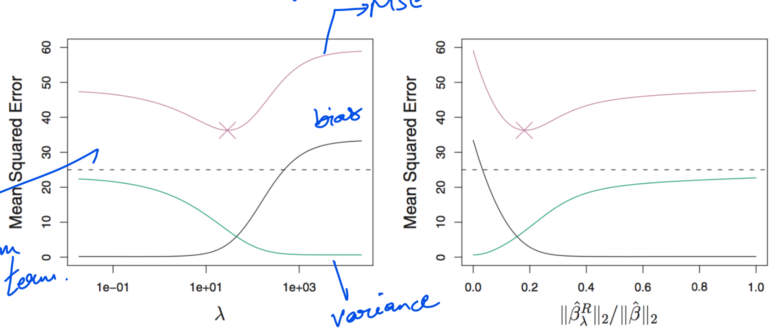


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Lasso encourages sparse models

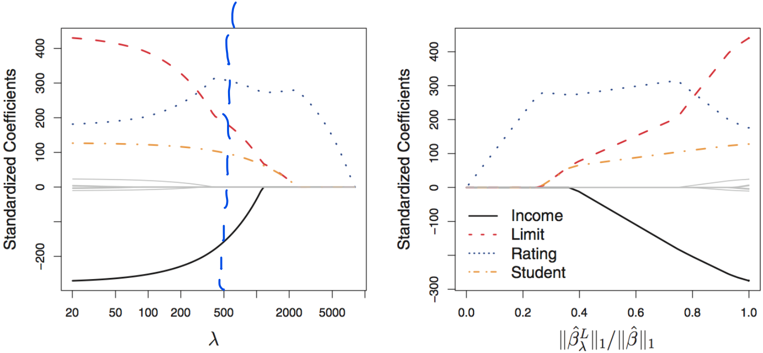


FIGURE 6.6. *The standardized lasso coefficients on the* Credit *data set are shown as a function of λ and $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.*

-Note $\|\beta\|_1 = |\beta_1| + \cdots |\beta_p|$

Another fomulation of ridge and lasso

- The Lasso's sparsity is better interpreted by an alternative formulation of Lasso and ridge regression $\lambda \uparrow$ then $s \downarrow$
- λ and s has some corresponding relationships.

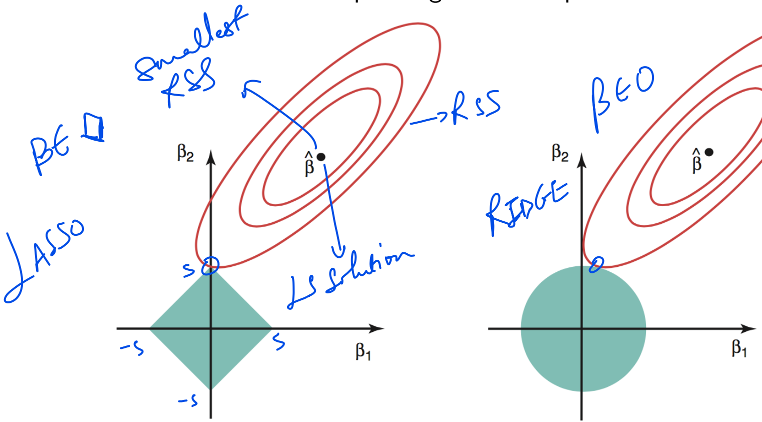


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \le s$ *and* $\beta_1^2 + \beta_2^2 \le s$*, while the red ellipses are the contours of the RSS.*

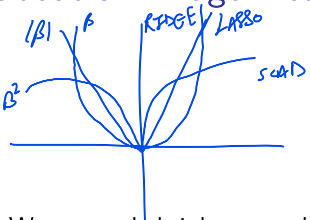
High-dimensional setting

- High-dimensional settings: the scenarios where the number of predictors p is comparable to or larger than the sample size n
- A situation common in modern biology and medical sciences, but less so in business
- Including more variables into the regression, we potentially might find some useful features, but this benefit needs to be weighted against including many noise features.
- Example: Suppose 20 features are useful to predict a numerical outcome, and we fix sample size (say at $n = 50$). Please compare the following three scenarios
  - i) Use all 20 features for prediction
  - ii) Use 18 of the above 20 features for prediction what if you eliminate 2 most important features.
  - iii) Use all 20 features, plus 100 noise features for prediction worst.
- Answer: i) is clearly better than iii). It is hard to compare i) and ii) only based on this abstract description.

High-dimensional setting

- High-dimensional settings: the scenarios where the number of predictors p is comparable to or larger than the sample size n
- A situation common in modern biology and medical sciences, but less so in business
- Including more variables into the regression, we potentially might find some useful features, but this benefit needs to be weighted against including many noise features.
- Example: Suppose 20 features are useful to predict a numerical outcome, and we fix sample size (say at $n = 50$). Please compare the following three scenarios
  - i) Use all 20 features for prediction
  - ii) Use 18 of the above 20 features for prediction
  - iii) Use all 20 features, plus 100 noise features for prediction
- Answer: i) is clearly better than iii). It is hard to compare i) and ii) only based on this abstract description.

More about shrinkage methods



- We covered shrinkage methods: ridge regression and Lasso
- Note that shrinkage methods are NOT limited these methods
- Other common ones include SCAD, elastic net, etc.
- In the so-called ultra high-dimensional settings (i.e., $p \gg n$), people sometimes use a two-step approach: marginal screening + shrinkage methods

$p = 1{,}000{,}000$
$n = 500$

$p = 1{,}000{,}000 \xrightarrow{MS} 1000 \rightarrow$ shrinkage

y

x_1 $|corr(y, x_1)|$
x_2 $|corr(y, x_2)|$
$\vdots$
$x_{1000000}$ $|corr(y, x_{1000000})|$