Section 9: Model regularization

STA 35C – Statistical Data Science III

Instructor: Akira Horiguchi

Fall Quarter 2025 (Sep 24 – Dec 12)
MWF, 12:10 PM – 1:00 PM, Olson 158
University of California, Davis

Overview

Based on Chapter 6 of ISL book James et al. (2021).

Section 7: Model selection and regularization

1. What goes wrong in high dimensions?

2. Shrinkage methods
   - Ridge regression
   - LASSO
   - Comparisons
   - Selecting the tuning parameter

What goes wrong in high dimensions?

When might $p \gg n$?

n is often limited due to cost, sample availability, or other considerations. Examples:

- Rather than predicting blood pressure on the basis of just age, sex, and BMI, one might also collect measurements for half a million single nucleotide polymorphisms (SNPs; these are individual DNA mutations that are relatively common in the population) for inclusion in the predictive model. Then $n \approx 200$ and $p \approx 500,000$.
- A marketing analyst interested in understanding online shopping patterns could treat as features all of the search terms entered by users of a search engine. This is sometimes known as the “bag-of-words” model. The same researcher might have access to the search histories of only a few hundred or a few thousand search engine users who have consented to share their information with the researcher. For a given user, each of the p search terms is scored present (0) or absent (1), creating a large binary feature vector. Then $n \approx 1,000$ and p is much larger.

Motivation

Linear regression models the relationship between the response and predictors as:

$$Y = \overbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p} + \epsilon. \tag{1}$$

Recall: the OLS estimator $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$ is the vector of parameters $(\beta_0, \beta_1, \ldots, \beta_p)^\top$ that minimizes the residual sum of squares (RSS)

$$\underline{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \text{where} \quad \hat{y}_i = \overbrace{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}} \tag{2}$$

- If the linear relationship is reasonable, the OLS estimates have low bias.
- If $n \gg p$, the OLS estimates tend to also have a low variance, and hence perform well on test observations.
- *If n is not much larger than p*, there can be a lot of variability in the least squares fit, resulting in overfitting/poor predictions. *If $p > n$*, there is no longer a unique OLS estimate, and the variance is infinite.
- *Try to reduce p?* Some predictors barely influence the response (if at all).
- *Feature selection*/*variable selection*, i.e. excluding irrelevant predictors from the linear model can lead to a more easily interpretable model.
- Least squares rarely ever yields any coefficient estimates that are exactly zero.

Here we explore fitting procedures other than least squares fit.

Alternative: Best subset selection

$\{a, b, c\}$

ϕ $\{b\}$ $\{c\}$ $\{bc\}$

$\{a\}$ $\{ab\}$ $\{ac\}$ $\{abc\}$

Given a positive integer $k \leq p$, what regression-coefficient estimators minimize the RSS *if at most k of the estimators are allowed to be non-zero?*

- *Subset selection:* identifies a subset of all p relevant predictors, then fits a model using least squares on the subset of predictors.
- For a long time, this problem was considered to be computationally intractable. How many possible subsets are there? 2^p $2^{10} = 1024$ $2^{20} \approx 1$ million
- Can approximate the solution using stepwise selection. (We did this in STA 35B.) $\approx p^2$ $10^2 = 100$ $20^2 = 400$
- 2016 paper[1]: “we demonstrate that our approach solves problems with n in the 1000s and p in the 100s in minutes to provable optimality”
- This is all we will say about this in STA 35C.

[1] https://projecteuclid.org/journals/annals-of-statistics/volume-44/issue-2/Best-subset-selection-via-a-modern-optimization-lens/10.1214/15-AOS1388.pdf

Shrinkage methods

Alternative: Shrinkage

Alternative: we can fit a model containing all p predictors using a technique that *constrains* or *regularizes* the coefficient estimates.

- I.e., a technique that shrinks the coefficient estimates towards zero (relative to the least squares estimate).
- This approach can also significantly reduce the variance of the coefficient estimates.

Two best-known techniques: *ridge regression* and *lasso (or LASSO)*.

Scaling predictor values

The OLS estimates are *scale equivariant*: multiplying values of predictors X_j by a constant c leads to scaling of the least squares coefficient estimates by a factor $\frac{1}{c}$.

- The ridge regression or LASSO coefficient estimates, however, can substantially change when multiplying predictors with a constant.
- Here it is better to work with predictor values which are all on the same scale. This can be achieved by *standardizing* the observations x_{ij} of predictors X_j:

$$\tilde{x}_{ij} = \frac{x_{ij}}{s_j}\,, \tag{3}$$

  where $s_j := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}$ is the standard deviation of the jth predictor, and where $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ is the sample mean of the observations of the jth predictor.

Remainder of slide deck will assume that predictors are standardized.

Shrinkage methods

Ridge regression

Ridge regression – Idea

For a user-chosen value $\lambda \geq 0$, the *ridge regression* coefficient estimates are defined as

Ridge Regression

$$\underbrace{\hat{\beta}^R_{0,\lambda}, \hat{\beta}^R_{1,\lambda}, \dots, \hat{\beta}^R_{p,\lambda}} := \operatorname*{arg\,min}_{\beta_0, \beta_1, \dots, \beta_p} \underbrace{\left\{ RSS + \overbrace{\lambda \sum_{j=1}^{p} \beta_j^2} \right\}}, \tag{4}$$

- Each value of λ results in a different set of ridge-regression coefficient estimates.

$p=1$: $RSS + \underbrace{\lambda \beta_1^2}$

gets smaller as $|\beta_1| \to 0$

What does adding the *shrinkage penalty* $\boxed{\lambda \sum_{j=1}^{p} \beta_j^2}$ do?

- It is small if $\beta_1, \dots, \beta_p$ are close to zero, and so the ridge regression estimates are shrunk toward zero compared to the OLS estimates.
- The tuning parameter λ controls the impact of the shrinkage penalty. The larger the λ, the stronger the shrinkage.
  - If $\lambda = 0$, then (4) is RSS, and we get back the OLS estimates (no shrinkage).
  - As $\lambda \to \infty$, RSS's impact on (4) becomes increasingly smaller. Thus minimizing (4) requires shrinking $\beta_1, \dots, \beta_p$ increasingly strongly.
- The shrinkage is only applied to $\beta_1, \dots, \beta_p$, but not to β_0, because we want to shrink the estimated association of each predictor X_j with the response Y.

Ridge regression – Illustration

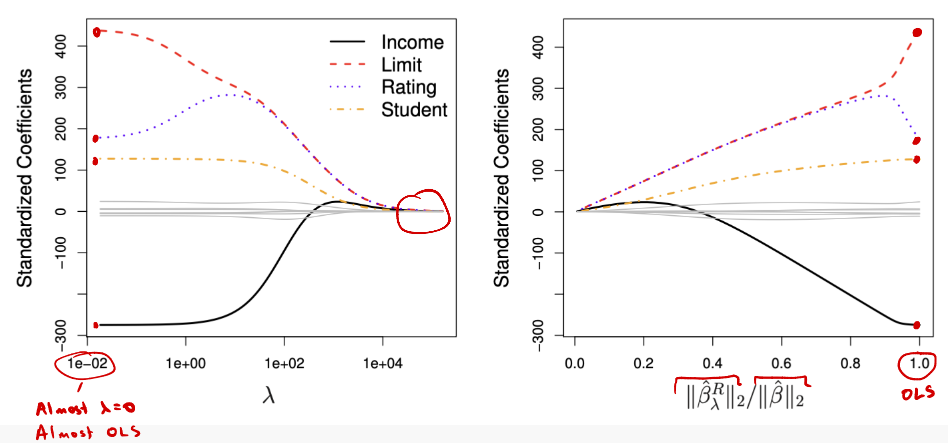


Figure 1: From James et al. (2021). Ridge regression coefficients (performed on standardized predictors) for the Credit data set in R.

Ridge regression – Advantages over least squares

1. For least squares (i.e., ridge regression with $\lambda = 0$), the variance is large but there is no bias. As λ increases, the flexibility of the ridge regression fit decreases, which reduces variance but increases bias.

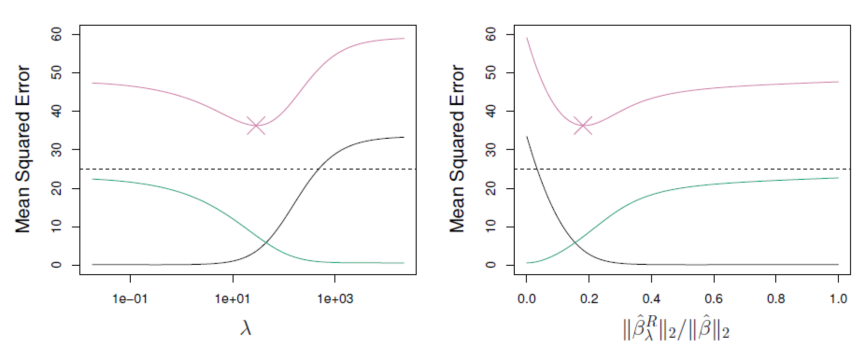


Figure 2: From James et al. (2021). Squared bias (in black), variance (in green), and test mean squared error (in purple) for the ridge regression predictions on a simulated data set. Horizontal dashed lines indicate the minimum possible MSE. Purple crosses indicate the ridge regression models for which the MSE is smallest.

Ridge regression works best when the OLS estimates have high variance.

Ridge regression – Advantages over least squares

2. Ridge regression has also a computational advantage over best subset selection, which requires searching through 2^p models.

Shrinkage methods

LASSO

Motivation

For any $0 < \lambda < \infty$, the ridge regression penalty shrinks cofficients towards zero, *but never exactly to zero.*

- Fine for prediction accuracy, but does not simplify model interpretation (an issue when the number of variables *p* is large).

LASSO

ridge: $\lambda \sum_{j=1}^{p} \beta_j^2$

For a user-chosen value $\lambda \geq 0$, the *LASSO* coefficient estimates are defined as

$$\hat{\beta}^L_{0,\lambda}, \hat{\beta}^L_{1,\lambda}, \ldots, \hat{\beta}^L_{p,\lambda} := \underset{\beta_0,\beta_1,\ldots,\beta_p}{\arg\min} \left\{ RSS + \overbrace{\lambda \sum_{j=1}^{p} |\beta_j|} \right\}. \tag{5}$$

- LASSO also shrinks the coefficients toward zero depending on λ, but *often forces some of the coefficients to be exactly zero.*
- Thus LASSO performs *variable selection*, and models generated from LASSO are usually easier to interpret than those produced by ridge regression.
- We say a model is *sparse* if many of the coefficients are exactly zero.

LASSO regression – Illustration

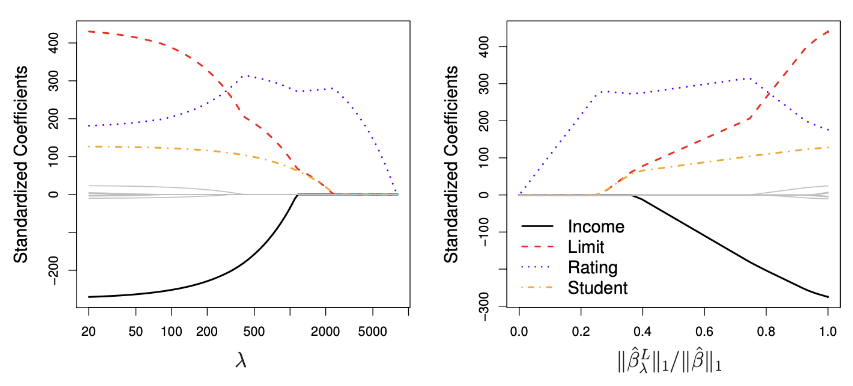


Figure 3: LASSO coefficients (performed on standardized predictors) for the Credit data set in R.

Why does LASSO perform variable selection? Alternate formulations

$x^2 + y^2 \le 1$

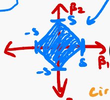
LASSO, ridge regression, and subset selection solve the following minimization problems for some $s \ge 0$:

$|\beta_1| + |\beta_2| \le s$

$p = 2$

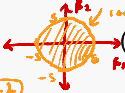
diamond

$$\text{LASSO:} \quad \min_{\beta} RSS \quad \text{subject to} \quad \boxed{\sum_{j=1}^{p} |\beta_j| \le s,} \tag{6}$$



circle radius s

$$\text{ridge regression:} \quad \min_{\beta} RSS \quad \text{subject to} \quad \boxed{\sum_{j=1}^{p} \beta_j^2 \le s^2;} \tag{7}$$



$\beta_1^2 + \beta_2^2 \le s^2$

$$\text{subset selection:} \quad \min_{\beta} RSS \quad \text{subject to} \quad \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}} \le s. \tag{8}$$

# of nonzero β_j's

- Helps explain why LASSO often performs variable selection, and why ridge regression never does.

Why does LASSO perform variable selection? Graphical comparison

Recall that RSS is a quadratic function of each β_j:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \text{where} \quad \hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{9}$$

$$\left(y_i - \left[\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right]\right)^2$$

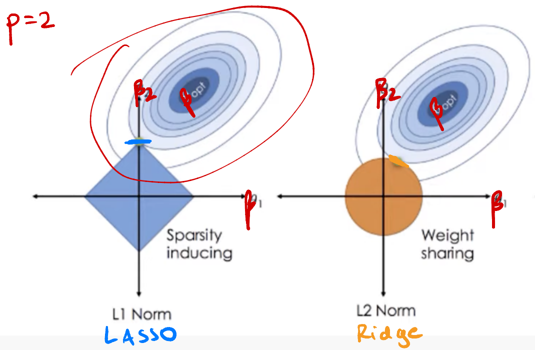


Figure 4: Here $p = 2$. The OLS solution is denoted as β_{opt}; ellipses are the contours of the RSS.
Left: Blue diamond is the constraint region $|\beta_1| + |\beta_2| \leq s$ for LASSO.
Right: Orange circle is the constraint region $\beta_1^2 + \beta_2^2 \leq s$ for ridge regression.
Source: `https://www.youtube.com/watch?app=desktop&v=iJE2fZcNPlA`

Shrinkage methods

Comparisons

Performance of LASSO and ridge regression

Regarding prediction accuracy, neither method universally dominates the other.

- Ridge regression might be better if e.g. the regression function truly depends on all *p* predictors, or vice versa if e.g. the regression function truly depends on only a small subset of the *p* predictors.

Example: LASSO > ridge regression

Here the simulated data is generated from a signal f that *depends on 2 out of 45 predictors.* $p=45$

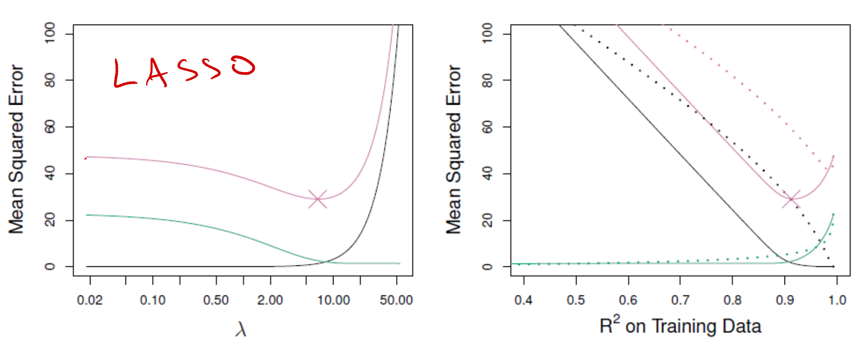


Figure 5: From James et al. (2021). Left: Plots of squared bias (in black), variance (in green), and test MSE (in purple) for LASSO on a simulated data set. Right: Comparison of squared bias, variance, and test MSE between LASSO (solid) and ridge (dotted). Both are plotted against their R^2 on the training data. The crosses in both plots indicate the LASSO model for which the MSE is smallest. (Plotting against R^2 can be used to compare models with different types of regularization.)

Example: Ridge regression > LASSO

p=45

Here the simulated data is generated from a signal f that *depends on all 45 predictors.*

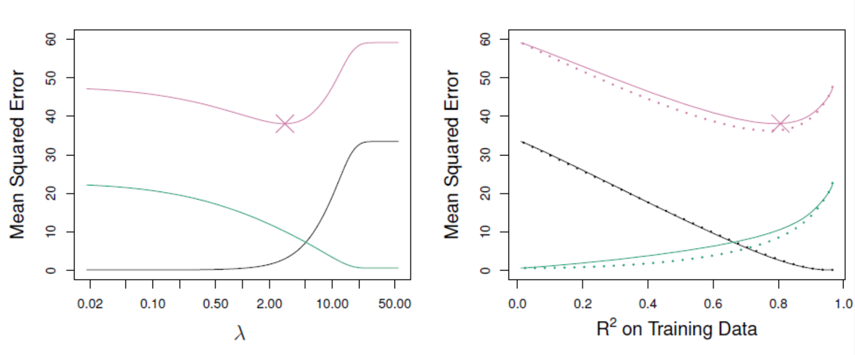


Figure 6: From James et al. (2021). Left: Plots of squared bias (in black), variance (in green), and test MSE (in purple) for LASSO on a simulated data set. Right: Comparison of squared bias, variance, and test MSE between LASSO (solid) and ridge (dotted). Both are plotted against their R^2 on the training data. The crosses in both plots indicate the LASSO model for which the MSE is smallest. (Plotting against R^2 can be used to compare models with different types of regularization.)

Special case

End of 11/5 lecture

Consider a special case:

- $n = p$;
- data matrix $\mathbf{X}$ is diagonal with 1's on the diagonal and 0's in all off-diagonal elements;
- we are performing regression without an intercept.

With these assumptions, the RSS is

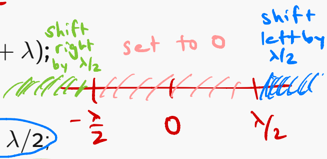
$$RSS = \sum_{j=1}^{p} \left(y_j - \beta_j\right)^2 \qquad (10)$$

and thus

- the OLS solution is given by $\hat{\beta}_j = y_j$;

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

- the ridge-regression solution is given by $\hat{\beta}_j^R = y_j/(1+\lambda)$;
- the LASSO solution is given by



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

Special case

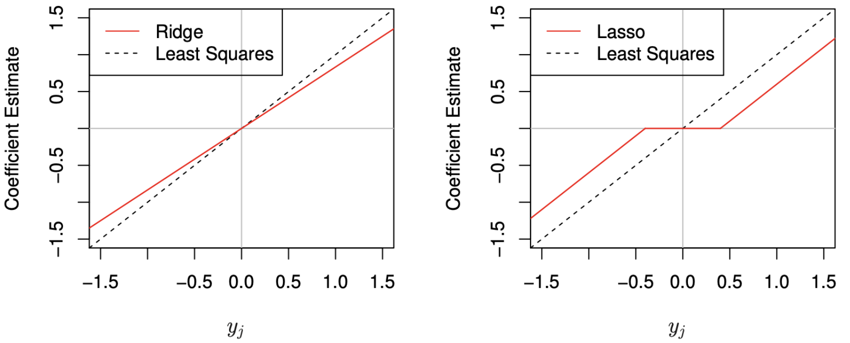


FIGURE 6.10. *The ridge regression and lasso coefficient estimates for a simple setting with $n = p$ and $\mathbf{X}$ a diagonal matrix with 1's on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: *The lasso coefficient estimates are soft-thresholded towards zero.*

Figure 7: From James et al. (2021).

Special case

Outside of this special case, the story is a little more complicated, but main ideas still hold approximately:

- ridge regression more-or-less shrinks every dimension of the data by the same proportion,
- LASSO more-or-less shrinks all coefficients toward zero by a similar amount, and sufficiently small coefficients are shrunken all the way to zero.

Shrinkage methods

Selecting the tuning parameter

Selecting the tuning parameter

Both LASSO and ridge regression require a parameter $\lambda \geq 0$, or equivalently a constraint $s \geq 0$ as described before, but how do we choose a value that might minimize the expected test MSE?

- Can use cross-validation: we choose a grid of λ values (as fine as possible), compute the CV error for each value of λ, and then select λ that minimizes the CV error.
- Can use R function `cv.glmnet()` – takes about as much time as a single OLS fit!

Coefficient estimates from cross-validation

Here the data comes from the Credit data set.

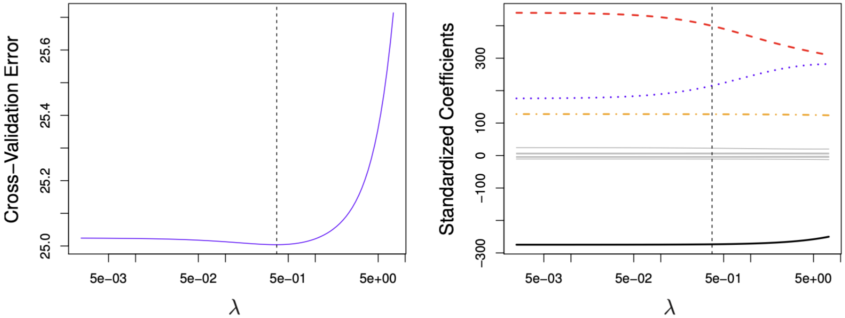


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various values of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

Figure 8: From James et al. (2021).

Coefficient estimates from cross-validation

Simulated data ($n = 50$); signal f *depends on 2 out of $p = 45$ predictors.*

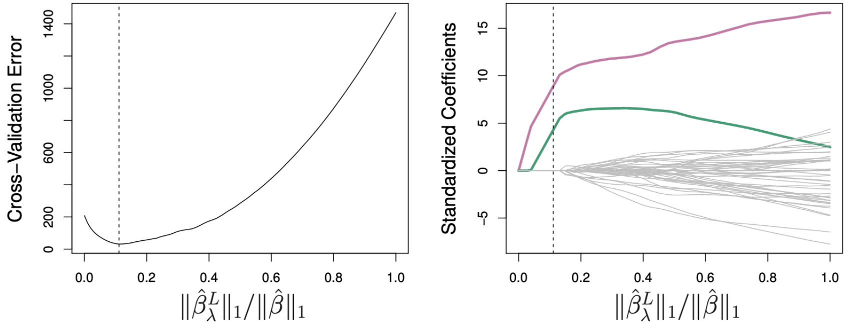


FIGURE 6.13. Left: *Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right: *The corresponding lasso coefficient estimates are displayed. The two signal variables are shown in color, and the noise variables are in gray. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

Figure 9: From James et al. (2021).