STATS 415: Shrinkage methods
Ridge regression and Lasso

Prof. Liza Levina

Department of Statistics, University of Michigan

Improving on Ordinary Least Squares

- Subset selection (of variables X)
- Shrinkage (of coefficients $\hat{\beta}$)
- Dimension reduction (of variables X) if p is large

Shrinkage

- Shrinking the estimated coefficients towards zero (in absolute value)
- Shrinkage reduces variance
- If some of the coefficients are shrunk to exactly zero, those variables can be removed.
- Advantages: efficient optimization methods exist; valid inference is being developed
- Disadvantages: shrinkage increases bias; does not always result in variable selection.
- We will cover two methods: ridge regression and the lasso

Ridge regression

- Ordinary least squares (OLS) estimates β's by minimizing

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

- Ridge regression estimates β's by minimizing

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

- $\lambda > 0$ is a tuning parameter to be determined

The ridge penalty

- The effect of this criterion is to add the ridge penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ to RSS, which is a goodness-of-fit measure
- To minimize RSS alone: set $\hat{\beta} = \hat{\beta}_{OLS}$
- To minimize the penalty alone: set $\hat{\beta} = 0$
- Adding them together has the effect of “shrinking” large values of β’s towards zero.
- The larger λ, the more shrinkage

The constrained optimization formulation

- Solving the ridge regression problem

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

  is mathematically equivalent to solving the constrained optimization problem

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s$$

- s and λ can be mapped to each other (one-to-one correspondence)

The ℓ_2 norm constraint

- The ℓ_2 norm of a vector β is defined as

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

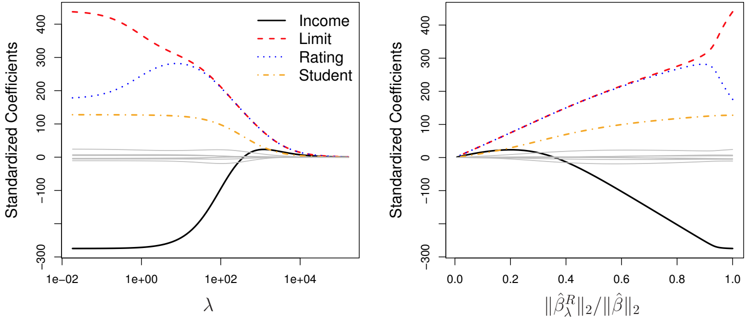
- Same as the usual vector norm in Euclidean space
- What shape does $\sum_{j=1}^{p} \beta_j^2 < s$ define?

Scaling in ridge regression

- In OLS, rescaling predictors rescales $\hat{\beta}$'s by the same constant: if we use $x_j' = cx_j$ instead of x_j, we'll get $\hat{\beta}_j' = \hat{\beta}_j/c$ instead of $\hat{\beta}_j$.
- t-statistics and p-values do not change with rescaling in OLS
- The ridge penalty term makes scaling much more important; we need different coefficients to be on the "same footing"
- Thus always standardize predictors before applying a shrinkage method

Credit card default data: ridge regression

- As λ increases, the coefficients shrink towards zero.
- An individual coefficient can go up or down
- Overall $\|\hat{\beta}_\lambda\|_2$ is always decreasing

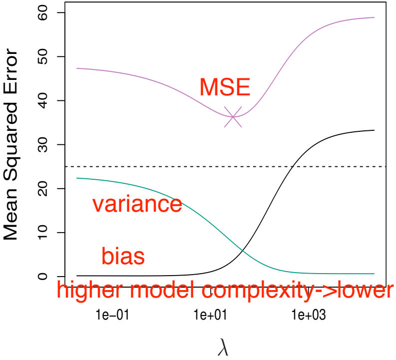


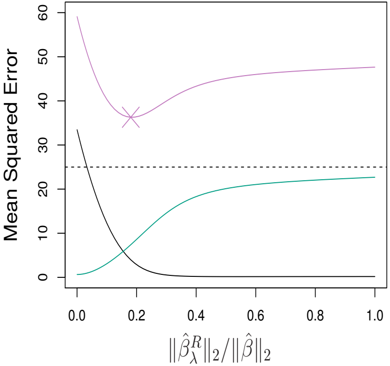
How can shrinking towards zero help?

- OLS estimates are unbiased if the model is true, but can be highly variable when
  - Predictors are highly correlated or collinear
  - There are outliers
  - n and p are of similar size or $n < p$
- The penalty increases bias but can substantially reduce variance.
- Need to choose the tuning parameter λ carefully to achieve the right bias/variance trade-off.

Ridge regression: Bias and variance

- Black: Bias^2; Green: Variance; Purple: MSE
- Increasing λ increases bias but decreases variance.
- Smaller λ corresponds to higher model complexity
- What does $\lambda = 0$ do? $\lambda = \infty$?





Computational advantages of ridge regression

computational advantage of ridge regression

- If p is large, then using the best subset selection approach requires searching through exponentially many possible models.
- Ridge regression is a quadratic optimization problem and can be solved in closed form
- For any given λ, there is a closed form solution
- Ridge regression works when $p > n$ and when predictors are collinear, both situations where OLS fails

The Lasso

- Ridge regression is not perfect: the final model still includes all variables (no selection means harder to interpret)
- A more modern alternative is the Lasso (LASSO = Least Absolute Shrinkage and Selection Operator).
- The Lasso works similarly to ridge, by shrinking the coefficients towards 0 and reducing variance while introducing some bias
- The only difference between ridge and lasso is in the form of the penalty term.

The Lasso penalty

- Ridge regression minimizes

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

- The Lasso estimates the β's by minimizing

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

- Equivalently, we solve

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

The ℓ_1 norm constraint

- Lasso uses the ℓ_1 norm penalty:

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

- The ℓ_1 norm of a vector β is defined as

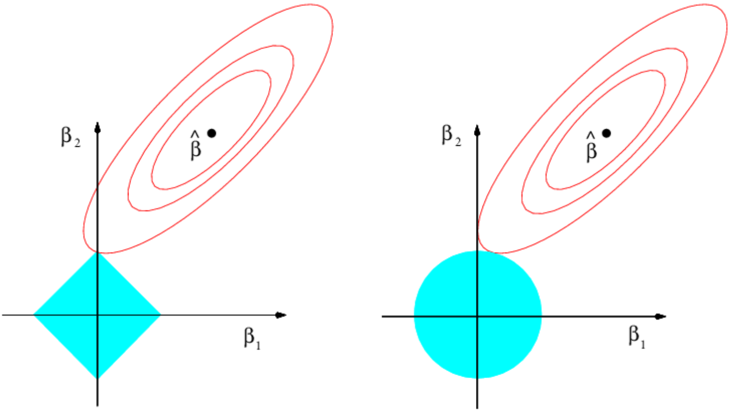
$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

- Sometimes referred to as the Manhattan distance
- What shape does $\sum_{j=1}^{p} |\beta_j| \leq s$ define?

Why does the penalty matter?

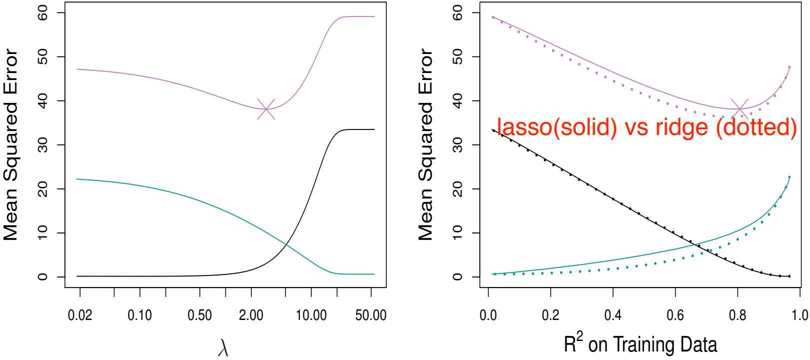
- Lasso seems like a very similar idea to ridge... but there is a big difference.
- Using this penalty, it could be proven mathematically that some coefficients will be set to exactly zero.
- Lasso can produce a model with good predictive power that is still simple to interpret.

Ridge vs lasso penalty illustration

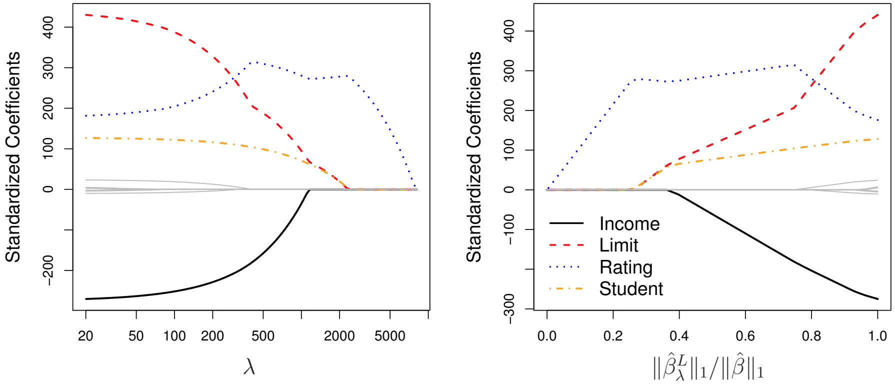


Lasso: Bias and variance

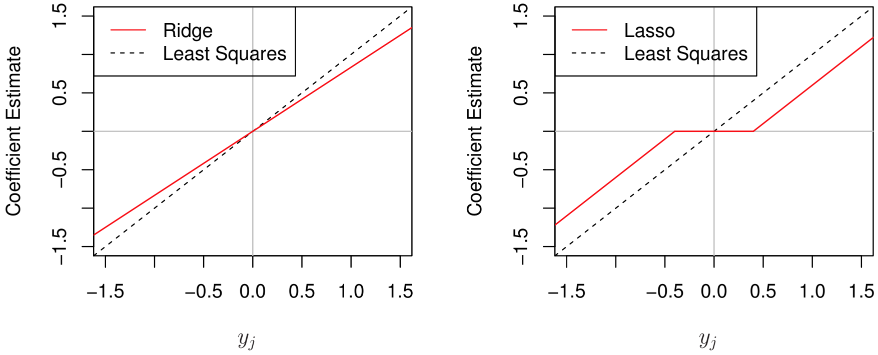
- Black: Bias^2; Green: Variance; Purple: MSE
- Increasing λ increases bias but decreases variance.
- Smaller λ corresponds to higher model complexity
- What does $\lambda = 0$ do? $\lambda = \infty$?
- Left: lasso; Right: lasso (solid) vs ridge (dotted)



Credit card default data: Lasso

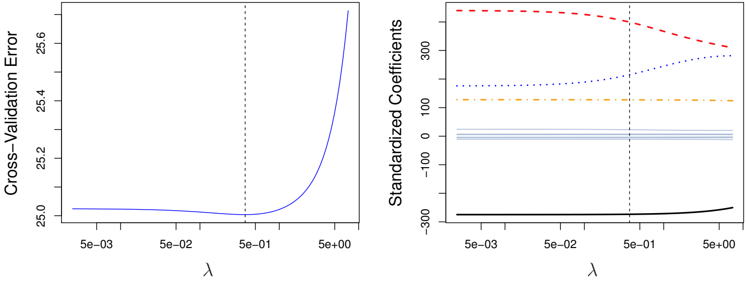


Ridge vs lasso coefficient shrinkage



Selecting the Tuning Parameter λ

- Choosing the right value of λ is crucial.
- Select a grid of potential values and select the value of λ that gives the smallest cross-validation error



Prediction vs interpretability

- Larger λ means more coefficients set to 0; but frequently better prediction is achieved with a smaller λ
- Sometimes a "one standard error" rule is used to select a more interpretable model: pick the largest λ such that the corresponding CV error is within one standard error (over cross-validation folds) of the best error
- Relaxed lasso: choose predictors with a larger λ, then refit the model with just those predictors without shrinkage (or little shrinkage)

Ridge and lasso: summary

- Both shrink coefficients towards 0 in order to reduce variance
- Both introduce bias
- Both work in cases when OLS fails, especially when $p > n$
- Both are computationally efficient
- Ridge does no variable selection and tends to do better at prediction
- Lasso shrinks some coefficients to 0 ("corners"), and thus does variable selection, but often predicts slightly less well than ridge