Lecture 3: Shrinkage Methods

UNIVERSITY OF SAN FRANCISCO

James D. Wilson

MATH 373

Plan for this Lecture

- Model Selection
- Regularization in Linear Regression
- Algorithms:
  - Ridge Regression (Tikhanov Regularization)
  - The Lasso
  - Elastic Net

Back to Linear Regression

Model:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

Estimate:

$$\hat{\beta}_{OLS} = \text{argmin}_{\beta} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right)$$

Questions:

1. What if we are primarily concerned with variable selection?
2. What if $p > n$? (high dimensional regression)

Recall: Best Subset Selection

Algorithm

Given: k predictors $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$

Loop: for (k in 1 to p)

1. Fit all $\binom{p}{k}$ models that contain k predictors
2. Pick M_k = the "best" among these models

Return: $M^* \in \{M_0, \ldots, M_p\}$: $M^* = \text{argmin}_j(S(M_j))$

- $S(M_j)$ = prediction criterion (Mallow's C_p, AIC, BIC, MSPE)

Recall: Best Subset Selection

Important Considerations:

1. Computational Complexity: must fit 2^p models
2. Algorithm is exhaustive: we *will find* the "best" model
3. Often replaced with approximate and less intensive algorithms:
    - Forward stepwise selection
    - Backward stepwise selection
    - Forward-backward stepwise selection

A Nice Alternative: Shrinkage Methods

1. Fits all p predictors using a technique that constrains or regularizes the coefficient estimates by optimizing a slightly different objective function
2. Equivalently, the techniques shrink coefficient estimates to zero
3. Variance of coefficient estimates are reduced as well! Particularly in high dimensional settings!

Ridge Regression

Model:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

Estimate:

$$\hat{\beta}_{Ridge} = \text{argmin}_{\beta} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

$\lambda \sum_{j=1}^{p} \beta_j^2$ acts as a shrinkage penalty to standard least squares regression since this value is small when β_j^2 is small.

Note: Also known as Tikhonov regularization

Ridge Regression: regularization perspective

Problem: The variance of $\hat{f}$ for OLS is often high $\Leftrightarrow$ predictions significantly change with small changes in X.

Reason: $X^T X$ is ill-conditioned $\Leftrightarrow$ either $p \approx n$ or variables suffer from multicollinearity:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$$

Solution: Ridge regression regularizes $(X^T X)$:

$$\hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T y$$

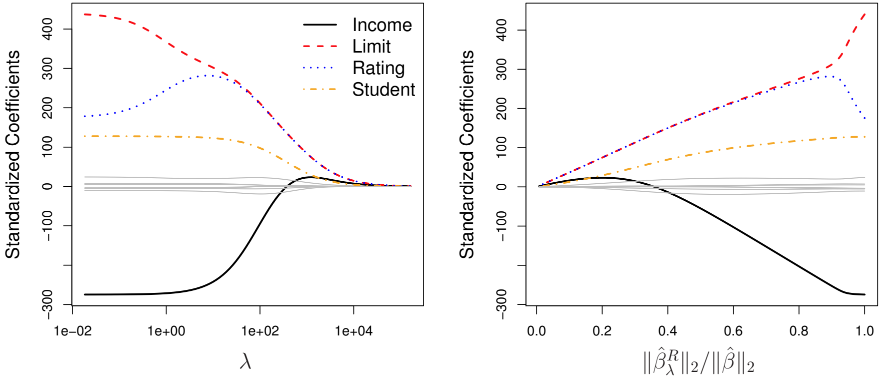
Ridge Regression

$$\hat{\beta}_{Ridge} = \text{argmin}_{\beta} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

λ: tuning parameter that adjusts the effect of the penalty

- $\lambda = 0 \quad \Rightarrow \quad \hat{\beta}_{OLS} = \hat{\beta}_{Ridge}$
- $\lambda \to \infty \quad \Rightarrow \quad \hat{\beta}_{Ridge} \to 0$
- λ chosen using cross validation: amazingly can computationally be determined for all possible values simultaneously!

Example: Comparison of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{Ridge}$



When does Ridge Regression outperform OLS?

1. **Computationally**: Ridge estimates for all values of λ can be determined simultaneously with one fit. Significant advantage over best subset selection that requires 2^p least squares fits.

2. **Model Accuracy**: OLS estimates often have high variance but low bias. Increases in λ lead to shrinkage, which subsequently leads to a major decrease in variance and only a slight increase in bias.

3. Key is to look across a grid of λ for best MSPE.

When does Ridge Regression outperform OLS?

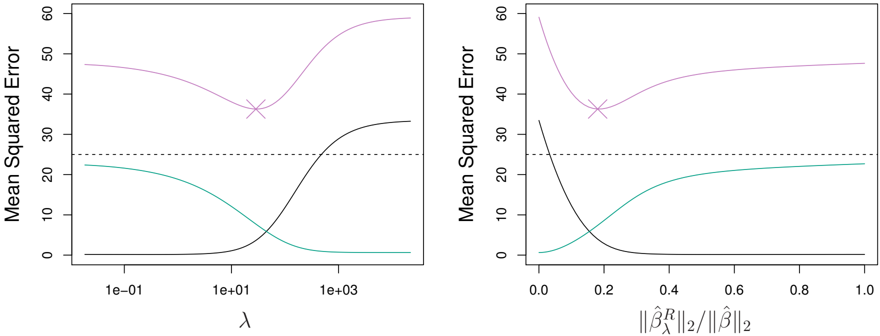


Figure: Squared bias (black), variance (green), and MSPE (purple) for $\hat{\beta}_{Ridge}$ on a simulated data set.

Weaknesses of Ridge Regression

- Requires a user - specified tuning parameter λ
- Interpretability of $\hat{\beta}_{Ridge}$
- **Subtle but important point**: The penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ shrinks β *towards* 0 but does not set any values *exactly* to 0.
  - **Exception**: $\lambda = \infty$ – here all β_j are exactly 0
  - **Consequence**: The saturated model is *always* chosen!

Question: Can we shrink some coefficients exactly to zero?

Least absolute shrinkage and selection operator

Model:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

Estimate:

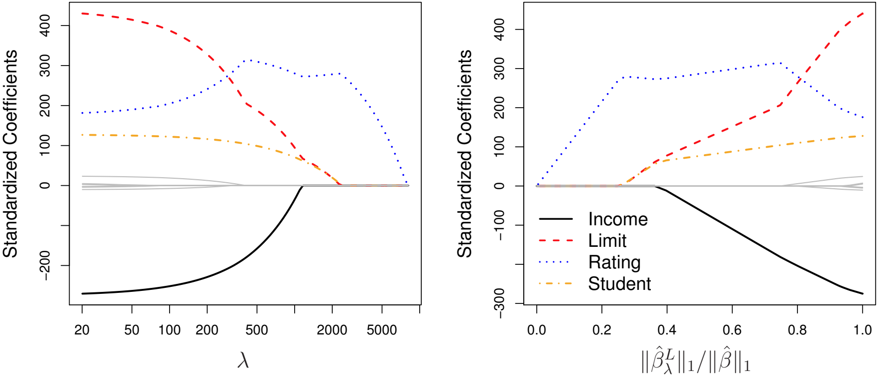
$$\hat{\beta}_{Lasso} = \text{argmin}_{\beta} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

$\lambda \sum_{j=1}^{p} |\beta_j|$ acts as a shrinkage penalty to standard least squares regression since this value is small when $|\beta_j|$ is small.

Historical Note on the Lasso

- From the paper "Regression shrinkage via the lasso" (1996) in *Journal of the Royal Statistical Society. Series B* by Robert Tibshirani (one of the authors of ISL and ESL)
- Considered by many to be the most influential modern statistical method
- Paper currently has 14243 citations! (as of October 27, 2015)
- Website:
  `http://statweb.stanford.edu/ tibs/lasso.html`

Variable Selection Property of Lasso



Note: Changing λ sets various subsets of β to 0! Why?

Re-formulations of Ridge and the Lasso

Both methods can be viewed as optimization problems.

- **Ridge Regression**:

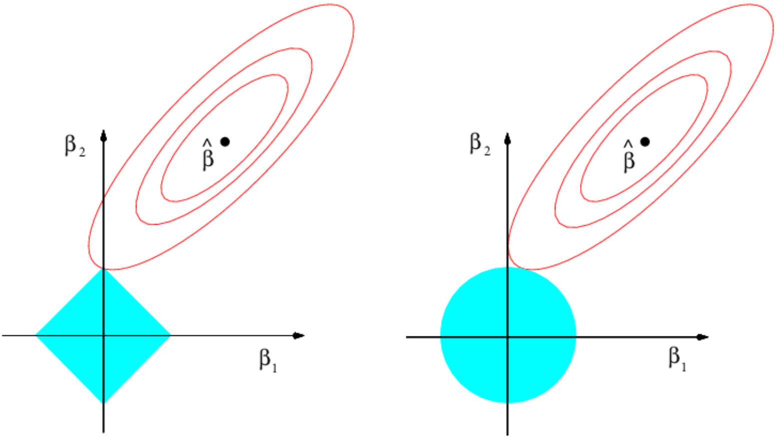
$$\text{minimize}_\beta \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right) \qquad \text{subject to} \qquad \sum_{j=1}^{p} \beta_j^2 \leq s$$

- **Lasso**:

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

Uh, ok so what? Explains the variable selection property of the Lasso!

Comparison of Lasso and Ridge



Often, the Lasso shrinks coefficients *exactly* to zero!

Comparison of Lasso and Ridge

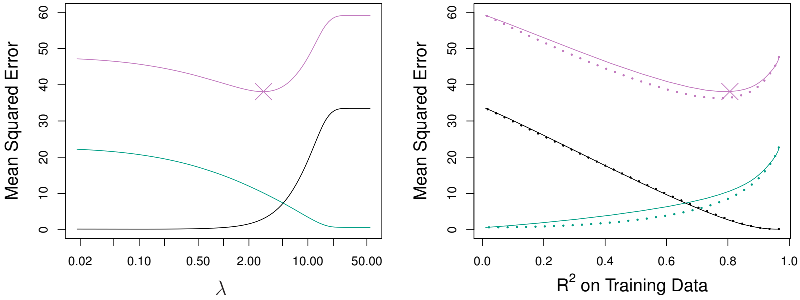


Figure: Squared bias (black), variance (green), and MSPE (purple). Dashed = Ridge, solid = Lasso

Note: Simulated data here included 45 / 45 non-zero coefficients. So, *no* variable selection is needed.

Comparison of Lasso and Ridge

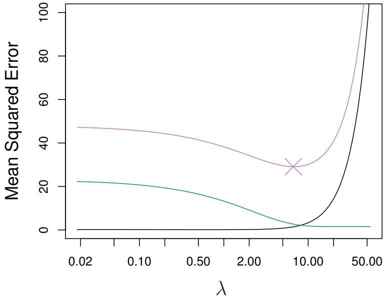


Figure: Squared bias (black), variance (green), and MSPE (purple). Dashed = Ridge, solid = Lasso

Note: Simulated data here included 2 / 45 non-zero coefficients. So, variable selection *is* needed.

A Bayesian Perspective of Ridge and Lasso

Let $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ be the parameters in a linear regression.

Bayesian Framework: Assume that β is a *random vector* with distribution $p(\beta)$. Here,

- $f(y|X, \beta)$ = likelihood of the data (Gaussian if ϵ is Gaussian)
- $p(\beta)$ = prior distribution of β
- $p(\beta|X, y)$ =posterior distrubtion of β given (X, y)

Bayes' Theorem gives us:

$$p(\beta|X, y) \propto f(y|X, \beta)p(\beta)$$

A Bayesian Perspective of Ridge and Lasso

Assumption 1: $p(\beta) = \prod_{i=1}^{p} g(\beta_i)$ (i.e. β_i's are iid).

Under Assumption 1, the regression model becomes:

$$y = \beta_0 + X_1\beta_1 + \ldots + X_p\beta_p + \epsilon$$

$$\beta_i \overset{iid}{\sim} g(x)$$

Properties

1. If $g(x)$ is a Gaussian distribution with $\mu = 0$, and $\sigma^2 = h(\lambda)$ then

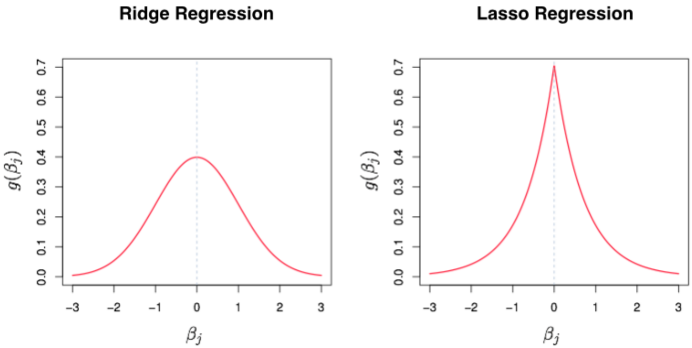
$$\hat{\beta}_{Ridge} = \text{Mode}(p(\beta|X, y))$$

2. If $g(x)$ is a Double Exponential distribution with $\mu = 0$, and $\sigma^2 = h(\lambda)$ then

$$\hat{\beta}_{Lasso} = \text{Mode}(p(\beta|X, y))$$

That is, by assuming a certain form of $g(x)$, we find that the Ridge and Lasso estimates are the maximum a posteriori (MAP) estimators for β.

A Bayesian Perspective



Another way of understanding the likelihood of shrinkage!

In general, the Lasso is best for variable selection / sparse relationships; Ridge for ill-conditioned problems.

Elastic Net: Combines Lasso and Ridge

Model:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

Estimate:

$$\hat{\beta}_{EN} = \text{argmin}_\beta \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \alpha\lambda \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

Best of both worlds? – well, this is more difficult to interpret!

Elastic Net Properties

1. $\alpha = 0$: reduces to Ridge Regression
2. $\alpha = 1$: reduces to Lasso
3. Has both properties of Ridge and Lasso:
    1. Reduces variance
    2. Variable selection
4. Recently proven that Elastic Net is equivalent to linear support vector machines.

Selecting λ

General Method: Grid search and cross-validation

1. Fix a value of λ
2. Estimate model and calculate average MSPE from k-fold cross-validation
3. Repeat the above procedure across a grid of λ
4. Choose λ that leads to smallest MSPE

Important: The above procedure can be done in parallel, easing computation.

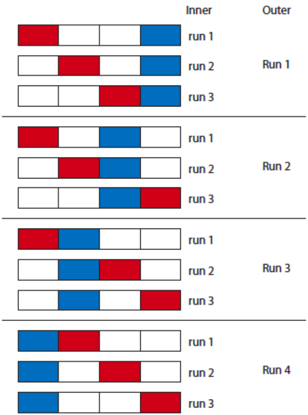
Selecting λ

Subtle yet Important Point:

1. Contrary to the fitting of a model in standard linear regression which relies upon minimizing MSE in training data only, we choose λ using cross-validation, which relies upon minimizing the MSPE of (cross-validation) test sets. Because of this, we typically hold out a test set initially and then run cross-validation on the training data.

2. Once λ is chosen, we then evaluate the MSPE on the original held-out test data.

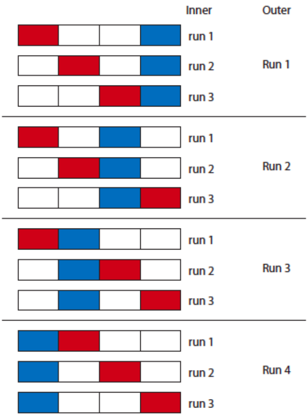
Selection λ



1) First choose a λ from **Validation set**: red (test) + white (training).

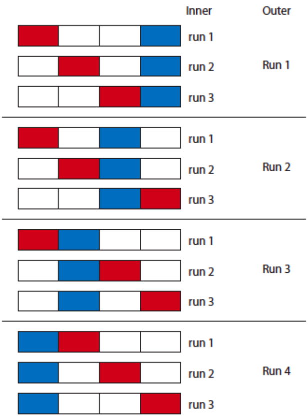
Here, we will get 4 of them $\lambda_1, \lambda_2, \lambda_3, \lambda_4$

Selection λ



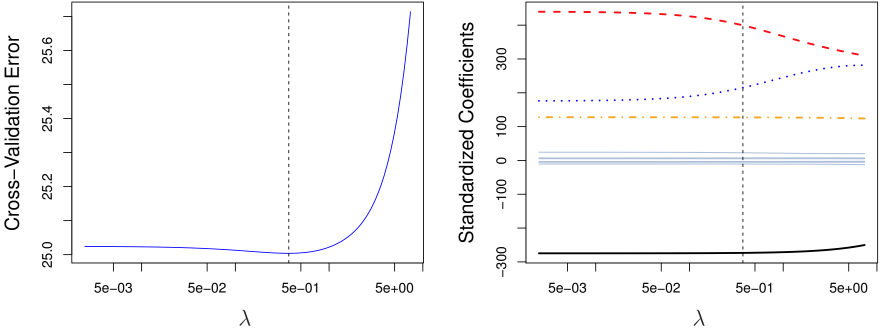
2) Then choose the best λ_i by i) training on **Validation set** and ii) testing on **Held out set**.

Selection λ

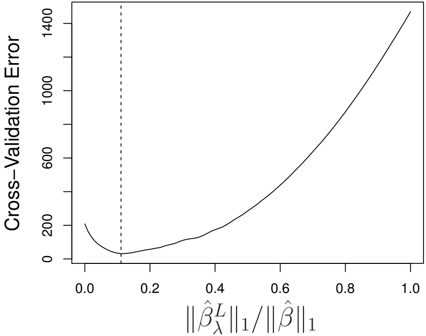


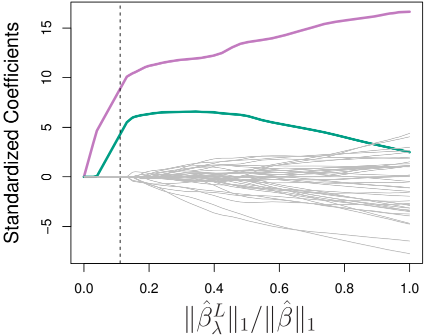
3) Fit final model using entire data set (**Validation** + **Held out**) with best λ_i

Selecting λ: Example with Ridge



Selecting λ: Example with Lasso





Implementing Shrinkage Methods with R

Next we will review how to implement the Lasso, Ridge Regression and Elastic Net in R.