FSAN/ELEG815: Statistical Learning

Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware

7. Lasso Regression

The ℓ_1 Norm and Sparsity

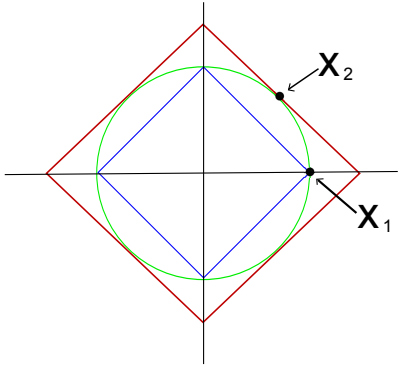
- The ℓ_0 norm is defined by: $\|x\|_0 = \#\{i : x(i) \neq 0\}$
  Sparsity of x is measured by its number of non-zero elements.
- The ℓ_1 norm is defined by: $\|x\|_1 = \sum_i |x(i)|$
  ℓ_1 norm has two key properties:
  - Robust data fitting
  - Sparsity inducing norm
- The ℓ_2 norm is defined by: $\|x\|_2 = (\sum_i |x(i)|^2)^{1/2}$
  ℓ_2 norm is not effective in measuring *sparsity* of x

Why ℓ_1 Norm Promotes Sparsity?

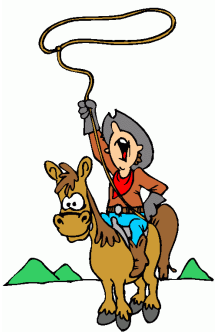
Given two N-dimensional signals:

- $x_1 = (1, 0, ..., 0) \rightarrow$ "Spike" signal
- $x_2 = (1/\sqrt{N}, 1/\sqrt{N}, ..., 1/\sqrt{N}) \rightarrow$ "Comb" signal

- x_1 and x_2 have the same ℓ_2 norm: $\|x_1\|_2 = 1$ and $\|x_2\|_2 = 1$.
- However, $\|x_1\|_1 = 1$ and $\|x_2\|_1 = \sqrt{N}$.



Least Absolute Shrinkage and Selection Operator (LASSO)



- LASSO combines shrinking of Ridge regression with variable selection. Tibshirani 1996.
- Difference between LASSO and Ridge regression is the penalty used

$$\hat{\mathbf{w}}^{ridge} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left[\sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{d} x_{ij} w_j)^2 + \lambda \sum_{j=1}^{d} w_j^2 \right]$$

$$\hat{\mathbf{w}}^{lasso} = \arg\min_{\mathbf{w}} \left[\sum_{i=1}^{N} (y_i - \sum_{j=1}^{d} x_{ij} w_j)^2 + \lambda \sum_{j=1}^{d} |w_j| \right]$$

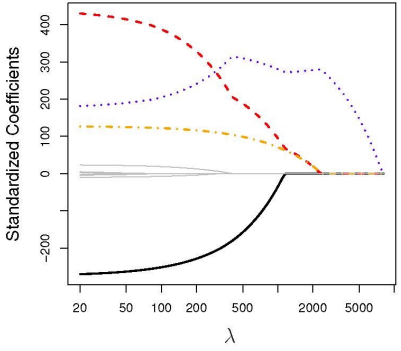
Least Absolute Shrinkage and Selection Operator (LASSO)

- LASSO coefficients are the solutions to the ℓ_1 optimization problem defined as

$$\begin{aligned}\hat{\mathbf{w}}^{lasso} &= \arg\min_{\mathbf{w}} \left[\sum_{i=1}^{N}(y_i - \sum_{j=1}^{d} x_{ij}w_j)^2 + \lambda \sum_{j=1}^{d}|w_j|\right] \\ &= \arg\min_{\mathbf{w}} \left[\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\mathbf{w})^2 + \lambda \sum_{j=1}^{d}|w_j|\right] \\ &= \arg\min_{\mathbf{w}} \left[(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w}) + \lambda||w_j||_1\right].\end{aligned}$$

- LASSO also shrinks the coefficients.
- ℓ_1 norm forces coefficients to zero when λ is large: **variable selection**.
- Lasso yields **sparse** models, keeping subset of variables.
- Unlike ridge regression, $\hat{\mathbf{w}}_\lambda^{lasso}$ has no closed form.

Lasso Regression Example Credit Data set



- Lasso performs better when a small number of predictors have strong coefficients, and the remaining predictors are small.
- Ridge regression performs better when the response is a function of many predictors.

The Variable Selection Property of the Lasso

One can show that the Ridge and Lasso regression coefficient estimates solve the following problems

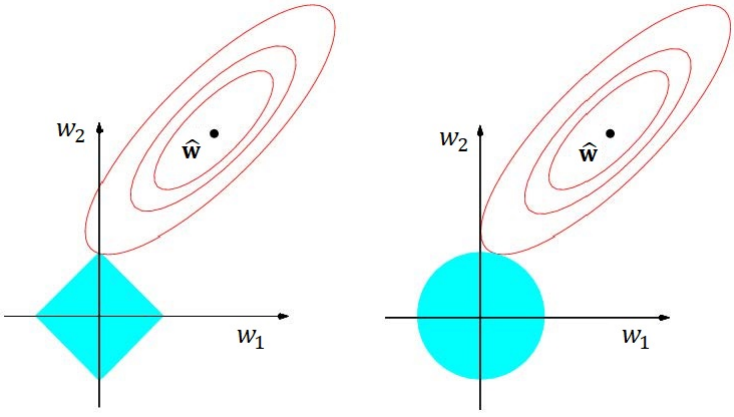
$$\hat{\mathbf{w}}^{ridge} = \underset{\mathbf{w}}{argmin}\{\sum_{i=1}^{N}(y_i - w_0 - \sum_{j=1}^{d} x_{ij}w_j)^2\} \quad (1)$$

$$\text{subject to } \sum_{j=1}^{d} w_j^2 \leq t$$

$$\hat{\mathbf{w}}^{lasso} = \underset{\mathbf{w}}{argmin}\{\sum_{i=1}^{N}(y_i - w_0 - \sum_{j=1}^{d} x_{ij}w_j)^2\} \quad (2)$$

$$\text{subject to } \sum_{j=1}^{d} |w_j| \leq t$$

The Variable Selection Property of the Lasso



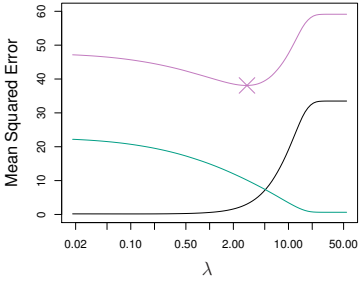
- RSS has elliptical contours, centered at the LS estimate.
- Constraint regions, $w_1^2 + w_2^2 \leq t$, and $|w_1| + |w_2| \leq t$.

Comparing the Lasso and Ridge Regression

The criteria to be analyzed for each case:

- Bias: Error that is introduced by approximating a real-life problem, by a much simpler model.
- Variance: Amount by which y would change is we estimated it using a different training data set.
- Training MSE: Mean squared error computed using the training data.
- Test MSE: Mean squared error computed using the test data.

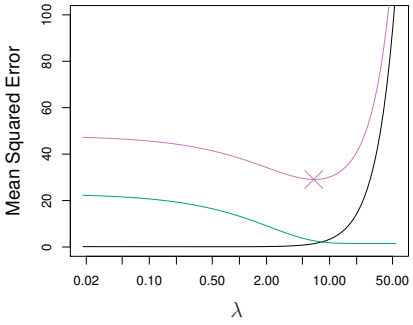
Comparing the Lasso and Ridge Regression



Simulated data set containing $d = 45$ predictors and $n = 50$ observations. For this figure all predictors were related to the response.

- Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso.
- Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed).

Comparing the Lasso and Ridge Regression

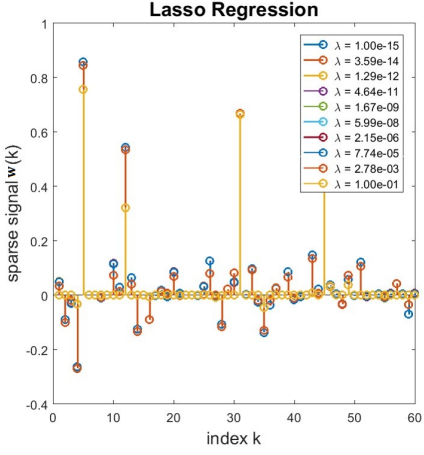


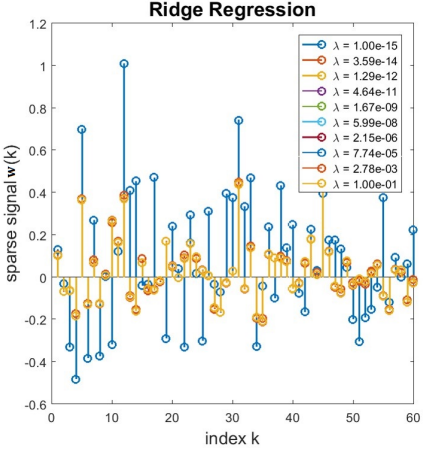
Here the the response is a function of only 2 out of 45 predictors.

- Left: Squared bias (black), variance (green), and test MSE (purple) for the lasso.
- Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed).

Lasso vs Ridge regression

- $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, where $\mathbf{X} \in \mathbb{R}^{40\times 60}$ is random Gaussian and $\boldsymbol{\epsilon}$ is noise.
- Original sparse signal is
  $w(k) = \delta(k-5) + 0.5\delta(k-12) + 0.9\delta(k-31) - 0.75\delta(k-45)$

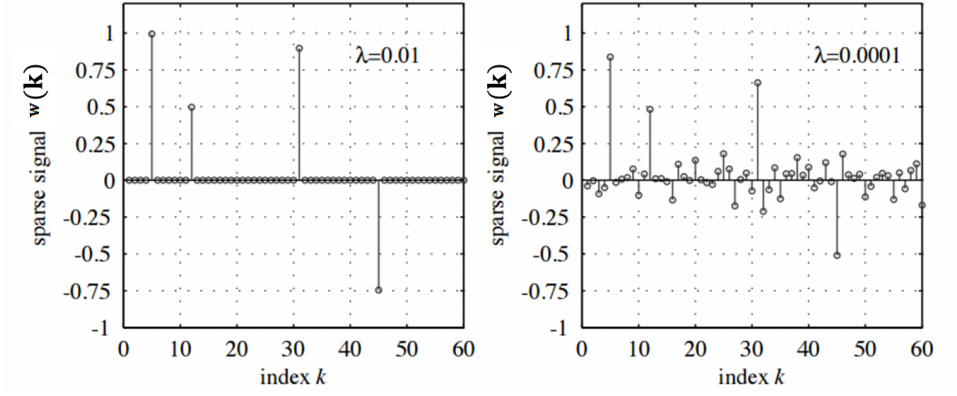




Example

$\mathbf{y} = \mathbf{X}\mathbf{w}$, where

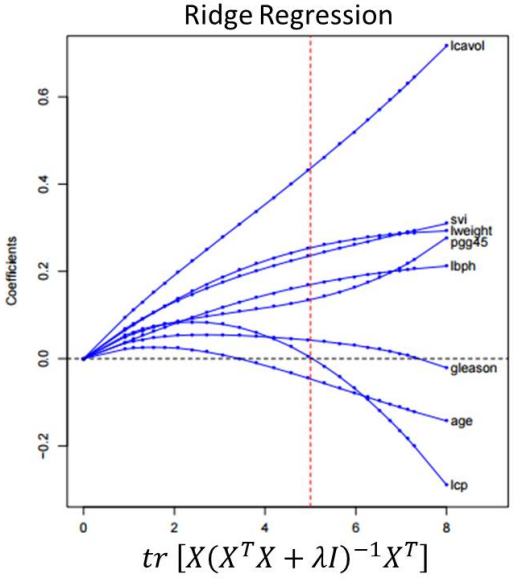
- $\mathbf{X}$ is a random Gaussian matrix $\in \mathbb{R}^{40\times 60}$.
- Original sparse signal is
  $w(k) = \delta(k-5) + 0.5\delta(k-12) + 0.9\delta(k-31) - 0.75\delta(k-45)$.
- The results for $\lambda = 0.01$ and $\lambda = 0.0001$ are presented

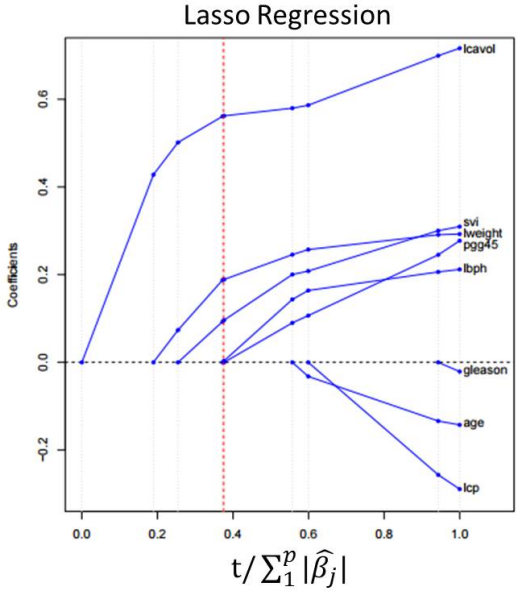


Example: Prostate Cancer

- Study by Stamey et al. (1989)
- Examines the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive radical prostatectomy.
- Variables: log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

Ridge vs Lasso Regression



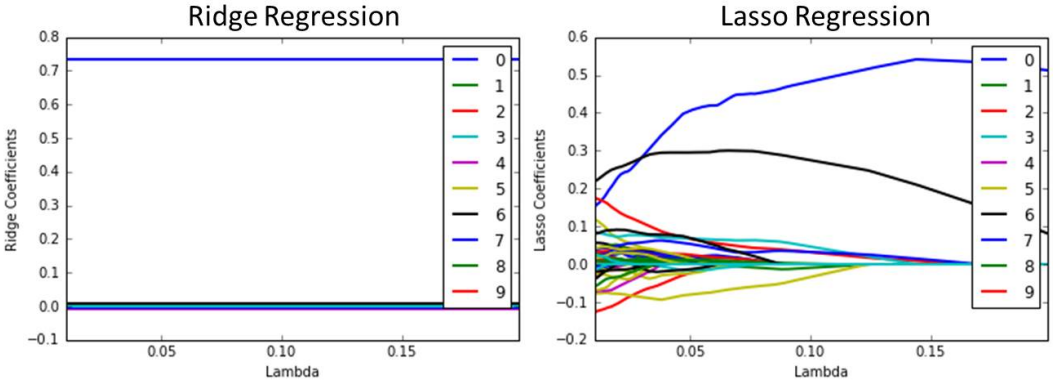


Example: Breast Cancer

We consider a classification problem involving a binary response variable $Y \in 0, 1$, describing the lymph node status of a cancer patient, and we have a covariate with $p = 7129$ gene expression measurements. There are $n = 49$ breast cancer tumor samples. The data is taken from West et al. (2001). It is known that this is a difficult, high noise classification problem.

Ridge vs Lasso Regression

Results for the 7129 predictors (Only first 10 labeled)



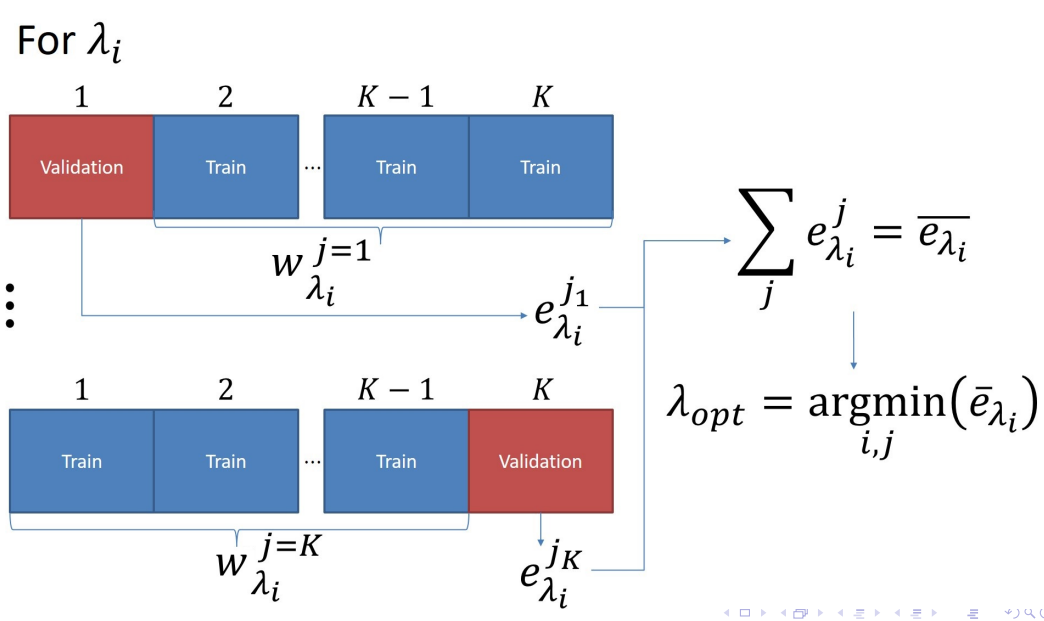
Choosing parameters: cross-validation

- Ridge and Lasso have regularization parameters.
- An *optimal* parameter needs to be chosen in a principled way

K- fold cross-validation: Split data into K equal (or almost equal) parts/folds at random.

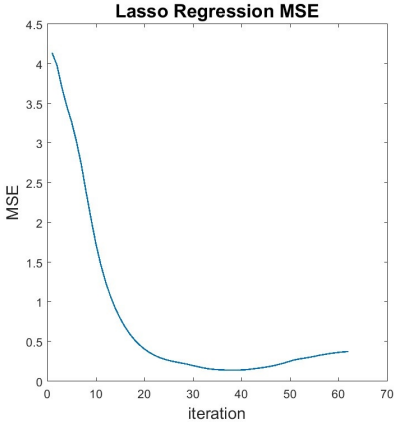
1: **for** each value λ_i **do**
2: **for** $j = 1, \cdots, K$ **do**
3: Fit model on data with fold j removed
4: Test model on remaining fold j^{th} test error
5: **end for**
6: Compute average test errors for parameter λ_i
7: **end for**
8: Pick parameter with smallest average error

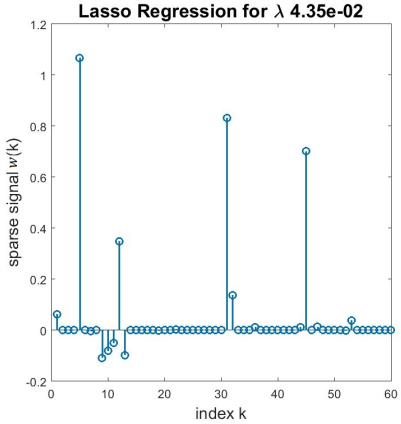
Choosing parameters: cross validation



Cross validation- Example K=5

- $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, where $\mathbf{X} \in \mathbb{R}^{40\times 60}$ is random Gaussian and $\boldsymbol{\epsilon}$ is noise.
- Original sparse signal is
  $w(k) = \delta(k-5) + 0.5\delta(k-12) + 0.9\delta(k-31) - 0.75\delta(k-45)$





Model selection vs Model assesment

- **Model selection:** estimate performance of different models in order to choose the “best" one
- **Model assessment:** having a chosen model, estimate its prediction error on new data
- When enough data is available, it is better to separate the data into three parts: train/validate, and test
- Typically: 50% train, 25 % validate, 25 % test.
- Test data is “kept in a vault", i.e. it is not used to fitting or choosing the model