FSAN/ELEG815: Statistical Learning

Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware

7: Lasso Regression

The l_2 Norm and Sparsity

- The l_0 norm is defined by: $\|\mathbf{x}\|_0 = \sharp\{i : x(i) \neq 0\}$
  Sparsity of **x** is measured by its number of non-zero elements

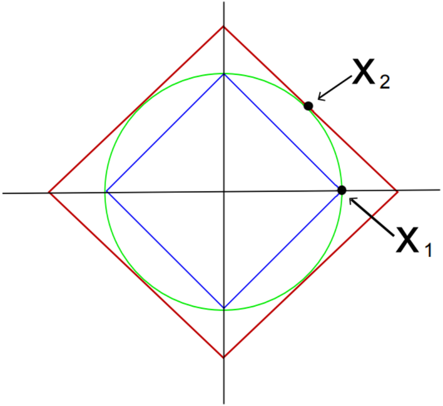
- The l_1 norm is defined by: $\|\mathbf{x}\|_1 = \sum_i |x(i)|$
  l_1 norm as two key properties:
  - Robust data fitting
  - Sparsity inducing norm

- The l_2 norm is defined by: $\|\mathbf{x}\|_2 = (\sum_i |x(i)|^2)^{1/2}$
  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



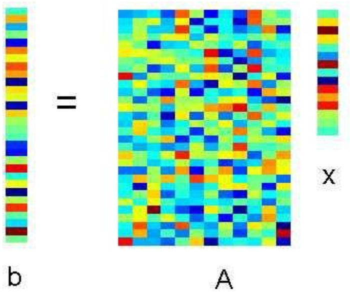
- x_1 and x_2 have the same ℓ_2 norm: $\|x_1\|_2 = 1$ and $\|x_2\|_2 = 1$.
- However, $\|x_1\|_1 = 1$ and $\|x_2\|_1 = \sqrt{N}$.

l_1 Norm in Regression

- Linear regression is widely used in science and engineering.

$$\text{Given} \quad \boldsymbol{A} \in R^{m\times n} \quad \text{and} \quad \boldsymbol{b} \in R^{m}; \;\; m > n$$

$$\text{Find} \quad \boldsymbol{x} \;\; \text{s.t.} \;\; \boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} \quad \text{(overdetermined)}$$



l_1 Norm Regression

Two approaches:

- Minimize the ℓ_2 norm of the residuals

$$\min_{\boldsymbol{x}\in R^n} || \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} ||_2$$

The ℓ_2 norm penalizes large residuals

- Minimizes the ℓ_1 norm of the residuals

$$\min_{\boldsymbol{x}\in R^n} || \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} ||_1$$

The ℓ_1 norm puts much more weight on small residuals

Matlab Code

$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2$

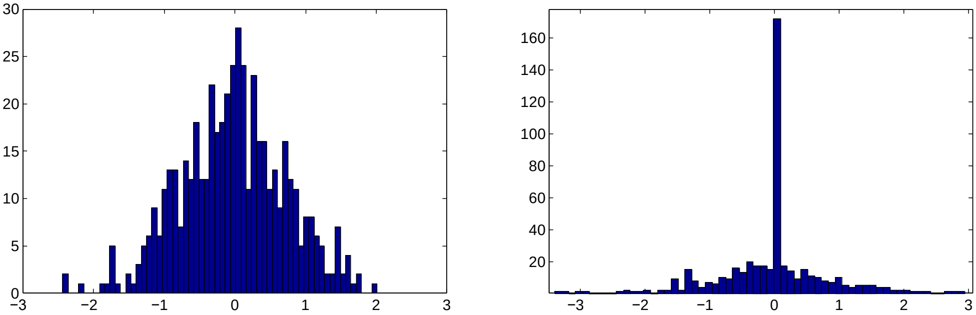
A=randn(500,150);
b=randn(500,1);
x=(A*A)$^{-1} * A * B$;   Least Squares Solution

$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_1$

A=randn(500,150);
b=randn(500,1);
X=medrec(b,A,max(A'*b),0,100,1e-5);

l_1 Norm Regression

$m = 500, n = 150$. $A = randn(m,n)$ and $b = randn(m,1)$

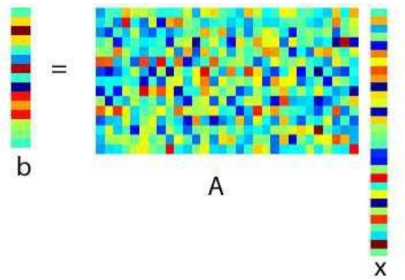


ℓ_2 Residuals $\qquad$ ℓ_1 Residuals

l_1 Norm Regression

Given $\boldsymbol{A} \in R^{m \times n}$ and $\boldsymbol{b} \in R^{m}$; $m < n$

Find $\boldsymbol{x}$ s.t. $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ (underdetermined)



l_1 Norm Regression

Two approaches:

- Minimize the ℓ_2 norm of $\boldsymbol{x}$

$$\min_{\boldsymbol{x}\in R^n} ||\boldsymbol{x}||_2 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

- Minimize the ℓ_1 norm of $\boldsymbol{x}$

$$\min_{\boldsymbol{x}\in R^n} ||\boldsymbol{x}||_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

Matlab Code

- $\min_{x \in R^n} ||\boldsymbol{x}||_2$ subject to $\boldsymbol{Ax} = \boldsymbol{b}$

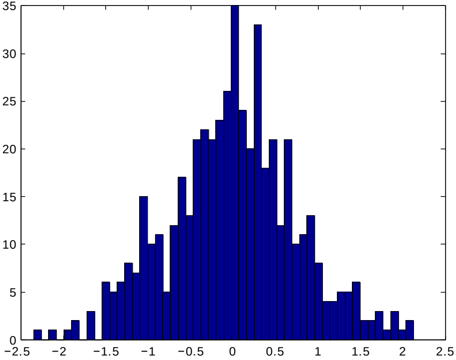
A=randn(150,500);
b=randn(150,1);
C=eye(150,500);
d=zeros(150,1);

X=lsqlin(C,d,[],[],A,b);

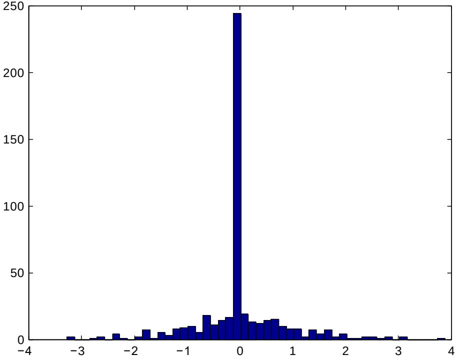
- In general:
  $\min_{x \in R^n} f(\boldsymbol{x})$ subject to $\boldsymbol{Ax} = \boldsymbol{b}$

*X= fmincon(@(**x**) f(**x**),zeros(500,1),[],[],A,b,[],[],options);*
where $f(\boldsymbol{x})$ is a convex function.

l_1 Norm Regression

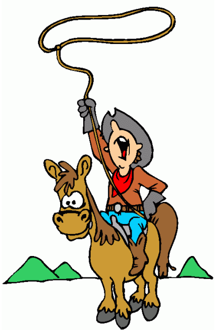


ℓ_2 Solution



ℓ_1 Solution

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

$$\hat{\mathbf{w}}^{\text{ridge}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min}\left[\sum_{i=1}^{N}(y_i - \sum_{j=0}^{d} x_{ij}w_j)^2 + \lambda\sum_{j=1}^{d} w_j^2\right]$$

$$\hat{\mathbf{w}}^{\text{lasso}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min}\left[\sum_{i=1}^{N}(y_i - \sum_{j=0}^{d} x_{ij}w_j)^2 + \lambda\sum_{j=1}^{d} |w_j|\right]$$

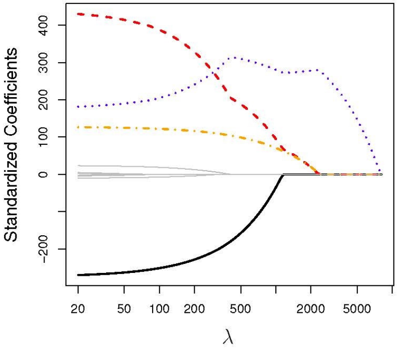
Least Absolute Shrinkage and Selection Operator (LASSO)

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

$$
\begin{aligned}
\hat{\mathbf{w}}^{\text{lasso}} &= \arg\min_{\mathbf{w}} \left[\sum_{i=1}^{N} (y_i - \sum_{j=1}^{d} x_{ij} w_j)^2 + \lambda \sum_{j=0}^{d} |w_j| \right] \\
&= \arg\min_{\mathbf{w}} \left[\sum_{i=1}^{N} (y_i - \mathbf{x}_i^T \mathbf{w})^2 + \lambda \sum_{j=0}^{d} |w_j| \right] \\
&= \arg\min_{\mathbf{w}} \left[(\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda ||\mathbf{w}||_1 \right].
\end{aligned}
$$

- 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

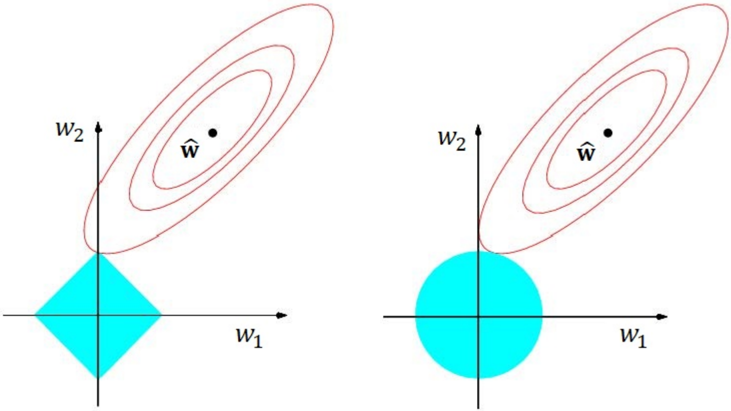
$$\hat{\mathbf{w}}^{\text{ridge}} = \arg\min_{\mathbf{w}}\{\sum_{i=1}^{N}(y_i - \sum_{j=0}^{d} x_{ij}w_j)^2\} \tag{1}$$

subject to $\sum_{j=0}^{d} w_j^2 \leq t$

$$\hat{\mathbf{w}}^{\text{lasso}} = \arg\min_{\mathbf{w}}\{\sum_{i=1}^{N}(y_i - \sum_{j=0}^{d} x_{ij}w_j)^2\} \tag{2}$$

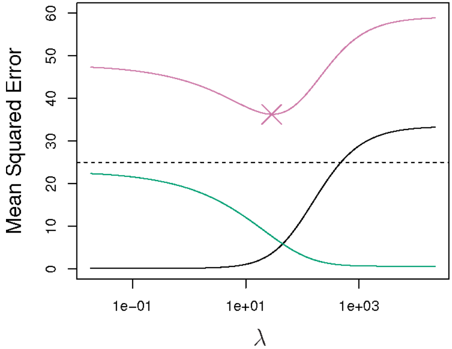
subject to $\sum_{j=0}^{d} |w_j| \leq t$

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 $|w_1| + |w_2| \leq t$.

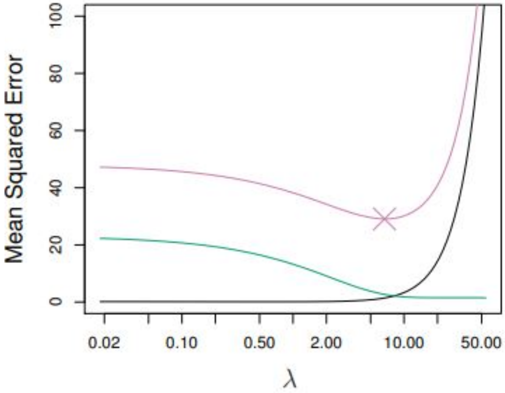
Comparing the Lasso and Ridge Regression



Simulated data set containing $d = 45$ predictors and $n = 50$ observations. Predictors related to the response.

- Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso.

Comparing the Lasso and Ridge Regression

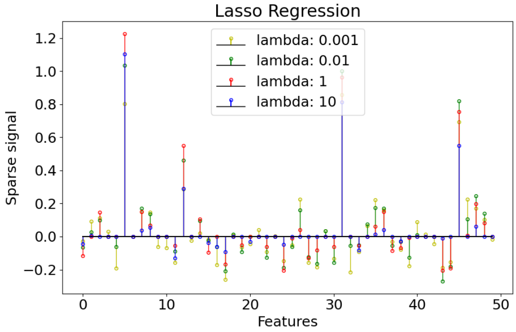


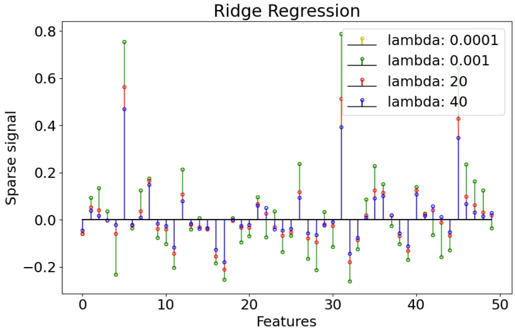
Here the response is a function of only 2 out of 45 predictors.

- Squared bias (black), variance (green), and test MSE (purple) for the lasso.

Lasso vs Ridge regression

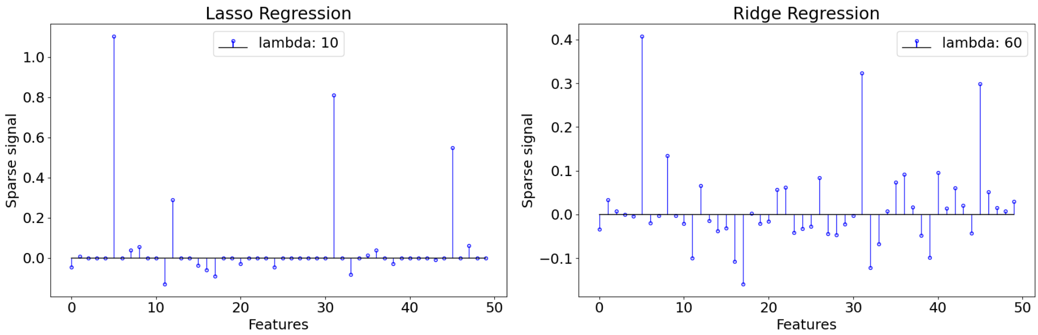
- $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, where $\mathbf{X} \in \mathbb{R}^{40\times 60}$ is random Gaussian and $\boldsymbol{\epsilon}$ is noise.
- Model given by
$w(k) = \delta(k-5) + 0.5\delta(k-12) + 0.9\delta(k-31) - 0.75\delta(k-45)$



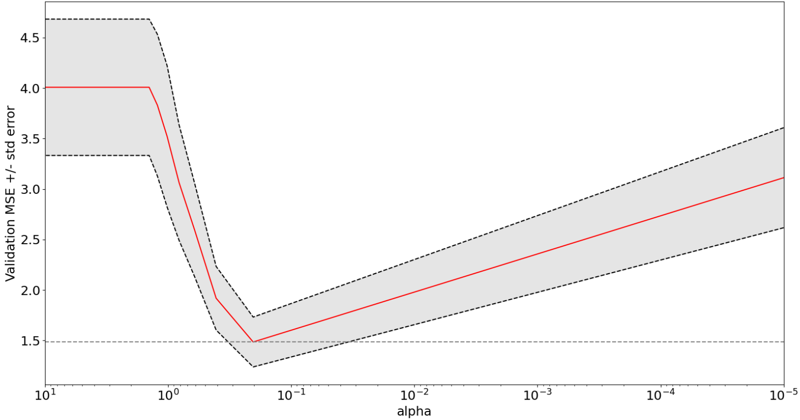


Lasso vs Ridge regression

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



Lasso hyperparameter optimization



Optimization of the alpha parameter through GridSearch with Cross-Validation and Mean Squared Error as the evaluation metric.

Iterative Calculation

- LASSO does not have a close form solution. Solved iteratively.
- Define $F(\mathbf{w}) = ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_1$.
- The solution to the LASSO problem is denoted as $\mathbf{w}_S$.
- Define an iterative procedure adding the non-negative term, having zero value at $\mathbf{w}_S$, $G(\mathbf{w}) = (\mathbf{w} - \mathbf{w}_S)^T(\alpha\mathbf{I} - \mathbf{X}^T\mathbf{X})(\mathbf{w} - \mathbf{w}_S)$, to the function $F(\mathbf{w})$.

The cost function is:

$$H(\mathbf{w}) = F(\mathbf{w}) + (\mathbf{w}-\mathbf{w}_S)^T(\alpha\mathbf{I}-\mathbf{X}^T\mathbf{X})(\mathbf{w}-\mathbf{w}_S), \tag{3}$$

where α is such that the added term is always nonnegative. It means $\alpha > \lambda_{max}$, where λ_{max} is the largest eigenvalue of $\mathbf{X}^T\mathbf{X}$.

$$\begin{aligned} H(\mathbf{w}) &= F(\mathbf{w}) + G(\mathbf{w}) \\ &= ||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_1 + (\mathbf{w}-\mathbf{w}_S)^T(\alpha\mathbf{I}-\mathbf{X}^T\mathbf{X})(\mathbf{w}-\mathbf{w}_S) \end{aligned}$$

Since $||\mathbf{w}||_1 = \mathbf{w}^T\text{sign}\{\mathbf{w}\} = ||\, w_1 sign(w_1), w_2 sign(w_2), \ldots, w_N sign(w_N)\,||_1$

$$\begin{aligned} H(\mathbf{w}) &= ||\mathbf{y}||_2^2 - \mathbf{w}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w}^T\text{sign}\{\mathbf{w}\} \\ &\quad + (\mathbf{w}-\mathbf{w}_S)^T(\alpha\mathbf{I}-\mathbf{X}^T\mathbf{X})(\mathbf{w}-\mathbf{w}_S) \end{aligned}$$

Iterative Calculation

$$
\begin{aligned}
H(\mathbf{w}) \quad = \quad & ||\mathbf{y}||_2^2 - \mathbf{w}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w}^T\text{sign}\{\mathbf{w}\} \\
& + (\mathbf{w}-\mathbf{w}_S)^T(\alpha\mathbf{I} - \mathbf{X}^T\mathbf{X})(\mathbf{w}-\mathbf{w}_S)
\end{aligned}
$$

Equating the gradient of $H(\mathbf{w})$ to zero:

$$
\begin{aligned}
\frac{\partial H(\mathbf{w})}{\partial \mathbf{w}^T} \quad &= \quad -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\text{sign}\{\mathbf{w}\} + 2(\alpha\mathbf{I} - \mathbf{X}^T\mathbf{X})(\mathbf{w}-\mathbf{w}_S) \\
0 \quad &= \quad -\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \frac{\lambda}{2}\text{sign}\{\mathbf{w}\} + \alpha\mathbf{w} - \mathbf{X}^T\mathbf{X}\mathbf{w} - (\alpha\mathbf{I} - \mathbf{X}^T\mathbf{X})\mathbf{w}_S \\
0 \quad &= \quad -\mathbf{X}^T\mathbf{y} + \frac{\lambda}{2}\text{sign}\{\mathbf{w}\} + \alpha\mathbf{w} - (\alpha\mathbf{I} - \mathbf{X}^T\mathbf{X})\mathbf{w}_S
\end{aligned}
$$

Rearranging the terms,

$$
\mathbf{w} + \frac{\lambda}{2\alpha}\text{sign}\{\mathbf{w}\} = \frac{1}{\alpha}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}_S) + \mathbf{w}_S
$$

Iterative Calculation

Corresponding iterative update

$$\mathbf{w}_{s+1} + \frac{\lambda}{2\alpha}\text{sign}\left\{\mathbf{w}_{s+1}\right\} = \frac{1}{\alpha}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}_s) + \mathbf{w}_s \tag{4}$$

How to solve it?

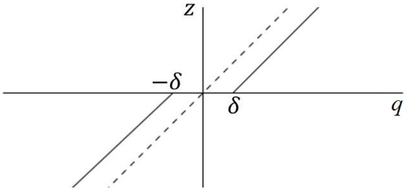
Note

The solution of the scalar equation $z + \delta \text{sign}(z) = q$, is obtained using soft-thresholding rule defined by a function soft(q, δ) as:

$$z = \text{soft}(q, \delta) = \begin{cases} q + \delta & \text{for} \quad q < -\delta \\ 0 & \text{for} \quad |q| \leq \delta \\ q - \delta & \text{for} \quad q > \delta \end{cases}$$

or

$$\text{soft}(q, \delta) = \text{sign}(q) \max\{0, |q| - \delta\}$$



Iterative Calculation

- The solution of $z + \delta \text{sign}(z) = q$ is $z = \text{soft}(q, \delta)$
- $\underbrace{\mathbf{w}_{s+1}}_{z} + \underbrace{\frac{\lambda}{2\alpha}}_{\delta} \text{sign}\left\{\underbrace{\mathbf{w}_{s+1}}_{z}\right\} = \underbrace{\frac{1}{\alpha}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}_s) + \mathbf{w}_s}_{q}$

Thus,

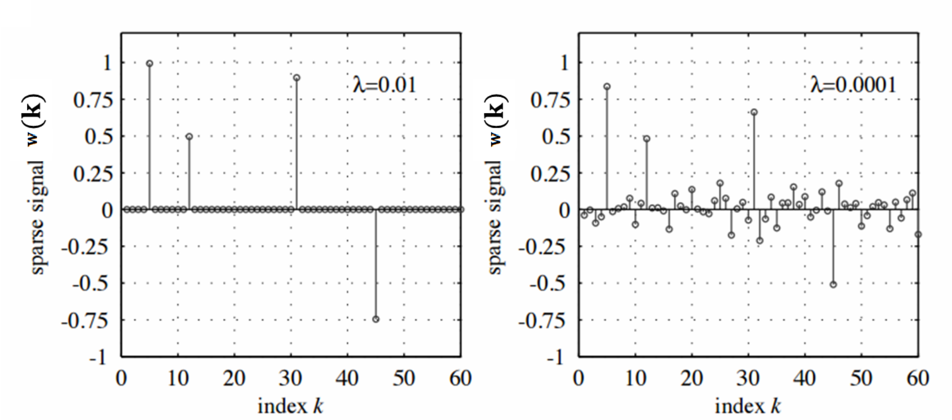
$$\mathbf{w}_{s+1} = \text{soft}\left(\frac{1}{\alpha}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}_s) + \mathbf{w}_s, \frac{\lambda}{2\alpha}\right) \tag{5}$$

This is the iterative soft-thresholding algorithm (ISTA) for LASSO minimization.

Example

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

- $\mathbf{X}$ is a random Gaussian matrix $\in \mathbb{R}^{40\times 60}$.
- Oracle model 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

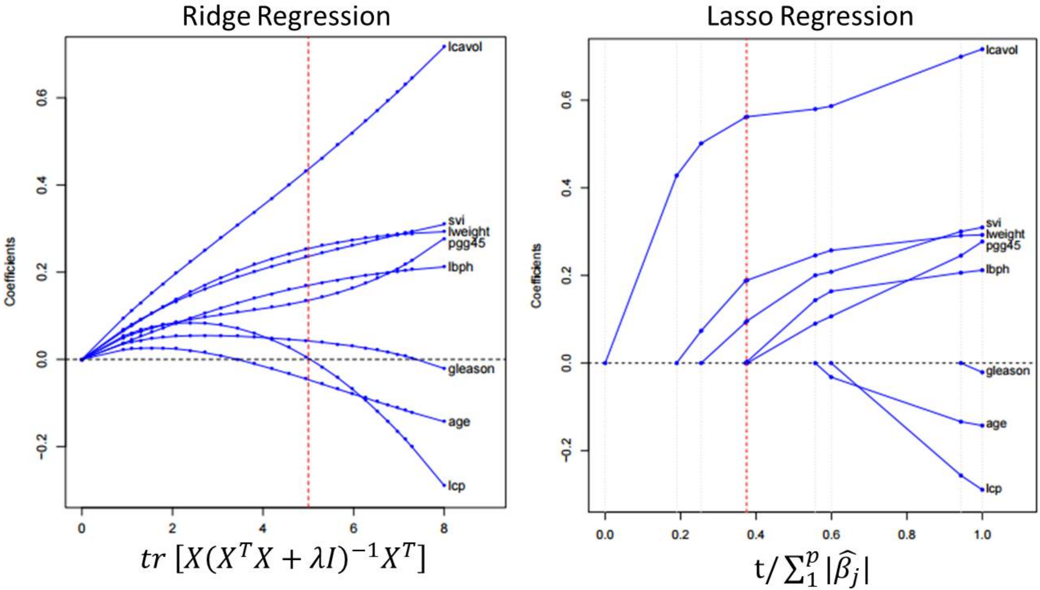


Example: Prostate Cancer

- 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.

Variable	Unit	Code
Cancer volume	log()	lcavol
Prostate weight	log()	lweight
age	-	age
Amount of benign prostatic hyperlasia	log()	lbph
Seminal Vesicle Invasion	-	svi
Gleason Score	-	gleason
Percentage of Gleason Score	4 or 5	pgg45

Ridge vs Lasso Regression



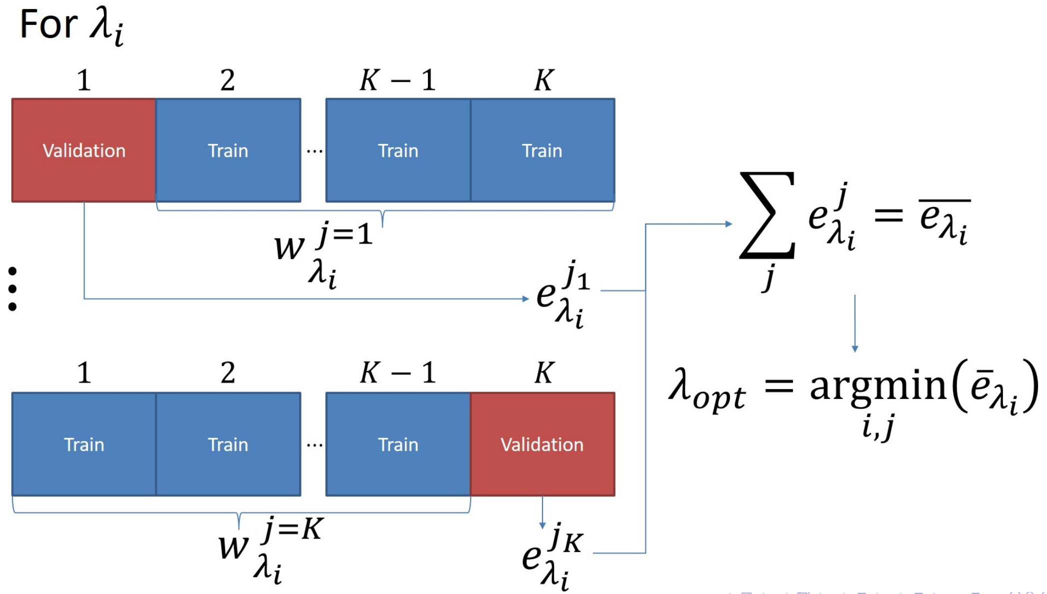
Choosing parameters: cross-validation

- 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.

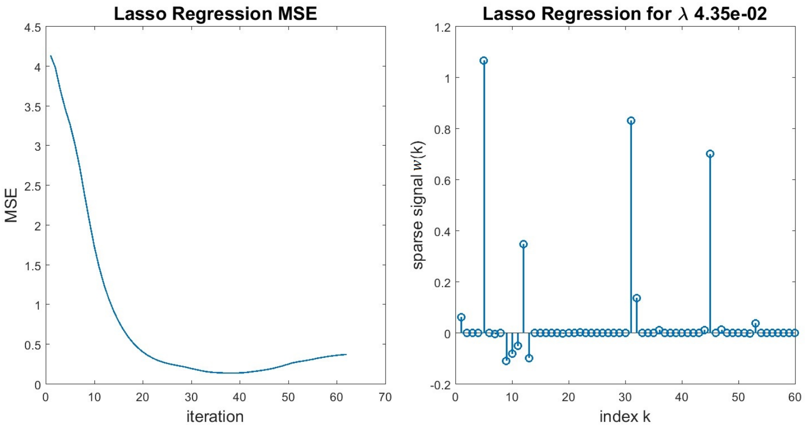
1: **for** each value λ_i **do**
2:     **for** $j = 1, \cdots, K$ **do**
3:         Fit model on data with fold j removed
4:         Test model on remaining fold j^{th} test error
5:     **end for**
6:     Compute average test errors for parameter λ_i
7: **end for**
8: Pick parameter with smallest average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