FSAN/ELEG815: Statistical Learning

Gonzalo 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: $\|\mathbf{x}\|_0 = \sharp\{i : x(i) \neq 0\}$
  The sparsity of $\mathbf{x}$ is measured by its number of non-zero elements

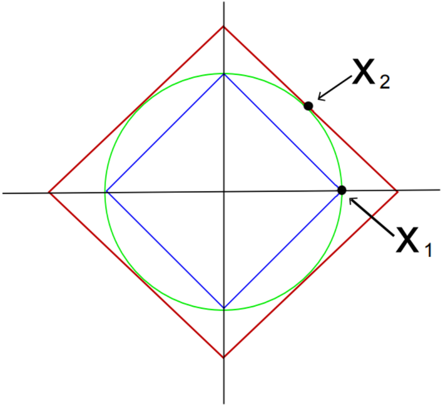
- The l_1 norm is defined by: $\|\mathbf{x}\|_1 = \sum_i |x(i)|$
  l_1 norm as two key properties:
  - Robust data fitting
  - Sparsity inducing norm

- 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 $\mathbf{x}$

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



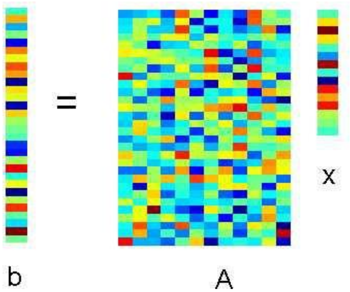
- x_1 and x_2 have the same ℓ_2 norm: $\|x_1\|_2 = 1$ and $\|x_2\|_2 = 1$.
- However, $\|x_1\|_1 = 1$ and $\|x_2\|_1 = \sqrt{N}$.

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

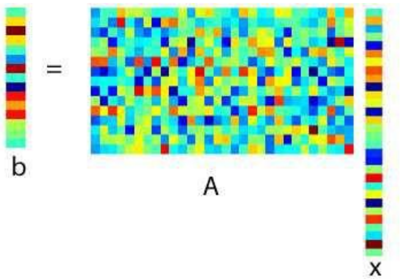
$$\min_{\boldsymbol{x}\in R^n} \| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \|_1$$

The ℓ_1 norm puts much more weight on small residuals

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 $\boldsymbol{x}$

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

Let's go to Python!

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

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

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

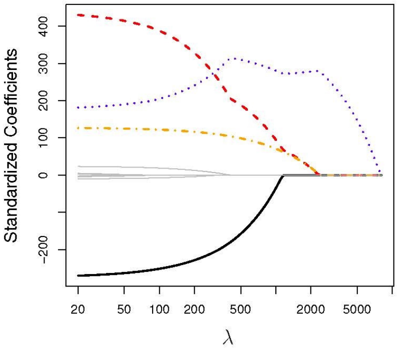
Least Absolute Shrinkage and Selection Operator (LASSO)

- 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

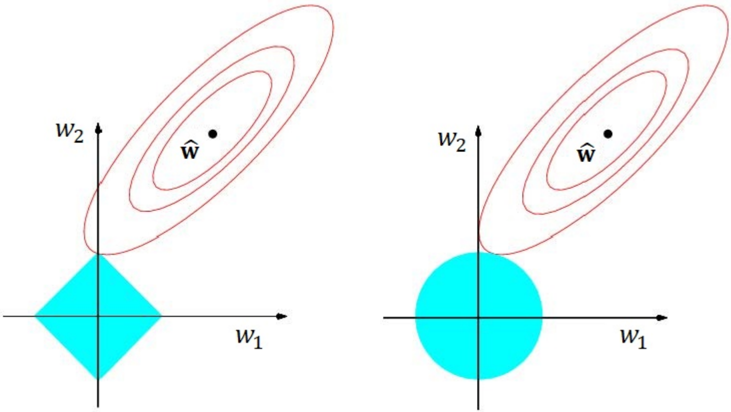
$$\hat{\mathbf{w}}^{\text{ridge}} = \arg\min_{\mathbf{w}} \{ \sum_{i=1}^{N} (y_i - \sum_{j=0}^{d} x_{ij} w_j)^2 \} \tag{1}$$

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

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

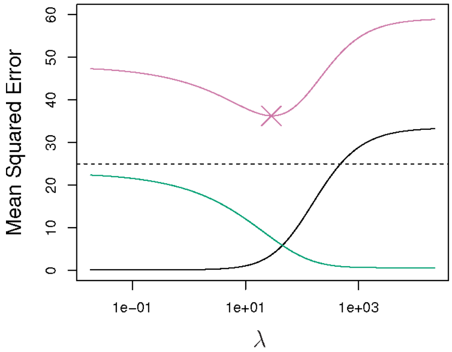
$$\text{subject to } \textstyle\sum_{j=0}^{d} |w_j| \leq t$$

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

Comparing the Lasso and Ridge Regression

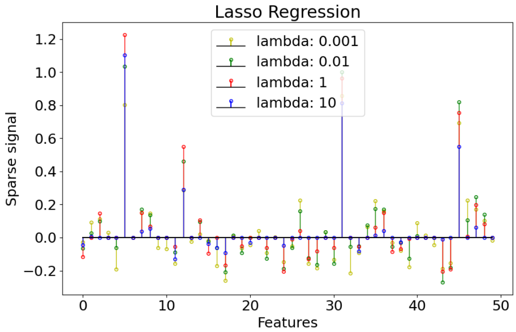


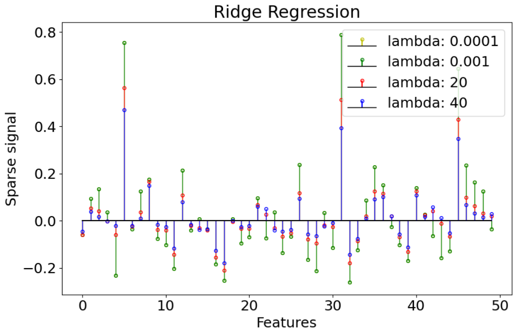
Simulated data set containing $d = 45$ predictors and $n = 50$ observations. Predictors related to the response.

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

Lasso vs Ridge regression

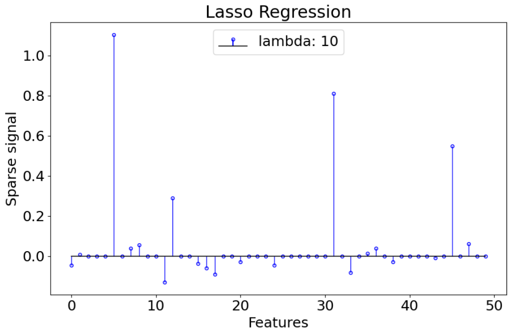
- $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, where $\mathbf{X} \in \mathbb{R}^{40\times 60}$ is random Gaussian and $\boldsymbol{\epsilon}$ is noise.
- Model given by
$$w(k) = \delta(k-5) + 0.5\delta(k-12) + 0.9\delta(k-31) - 0.75\delta(k-45)$$

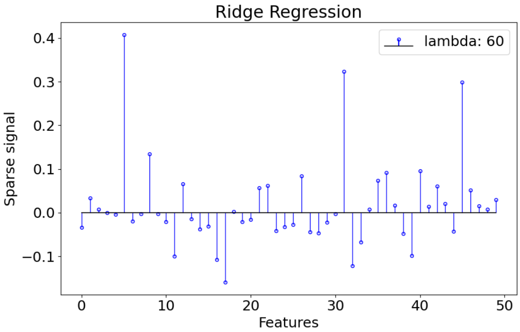




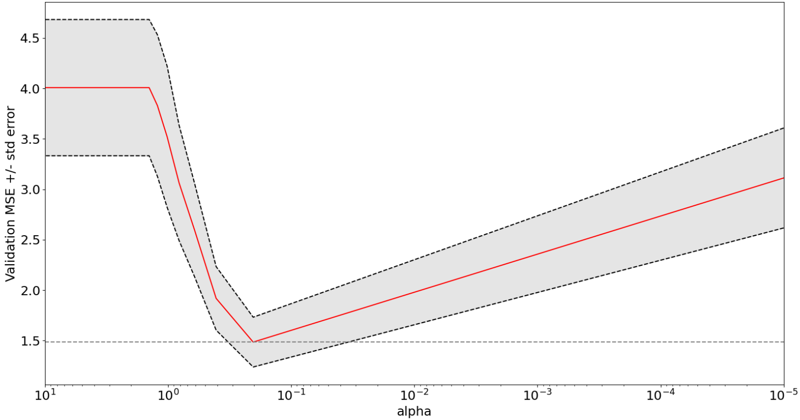
Lasso vs Ridge regression

- $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, 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

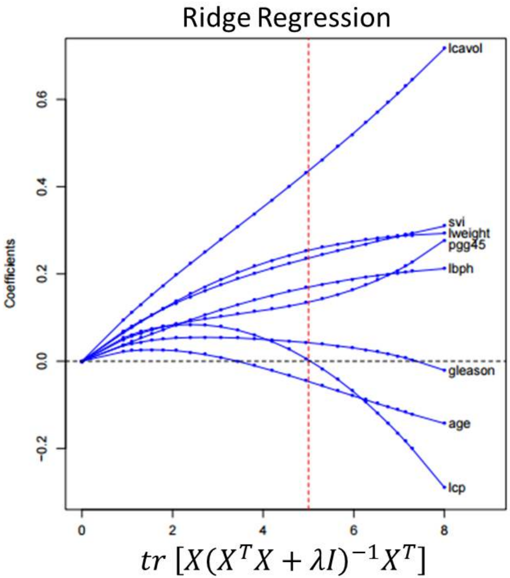
- LASSO does not have a closed-form solution. Solved iteratively:
  - Coordinate Descent Algorithm
  - Iterative Soft-Thresholding Algorithm (ISTA)

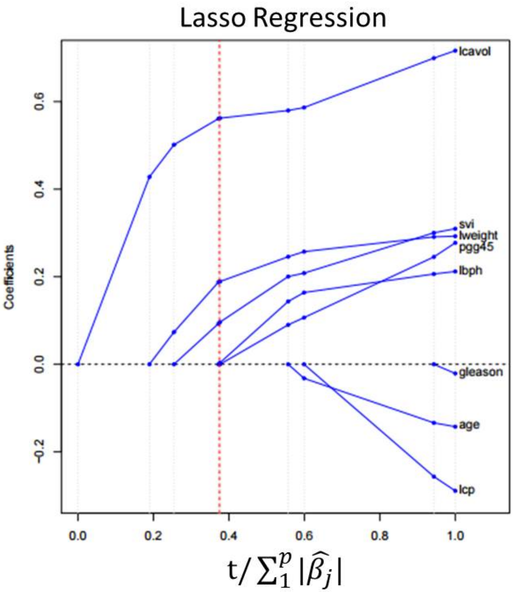
Example: Prostate Cancer

- 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 hyperplasia	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

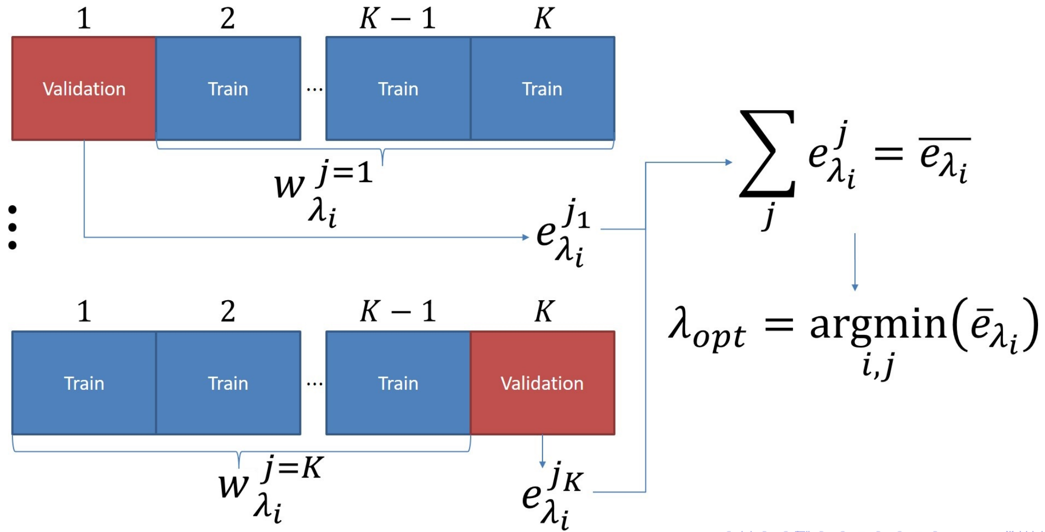
- Ridge and Lasso have regularization parameters.
- An *optimal* parameter needs to be chosen in a principled way

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

```
for each value λ_i do
    for j = 1, ··· , K do
        Fit model on data with fold j removed
        Test model on remaining fold j^th test error
    end for
    Compute average test errors for parameter λ_i
end for
Pick parameter with a smallest average error
```

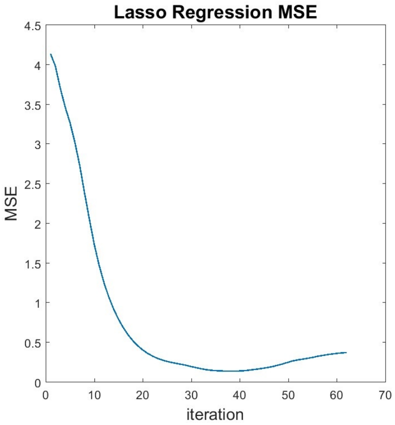
Choosing parameters: cross validation

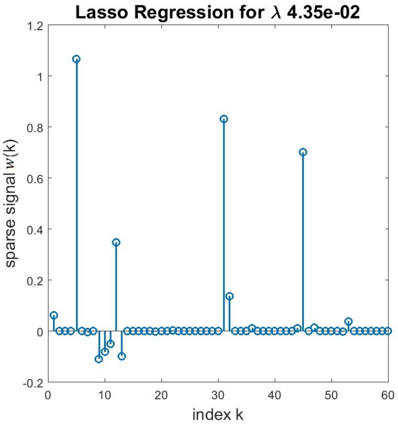
For λ_i



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 assessment

- **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 fit or choose the model