The Lasso

or: How I learned to Stop Worrying and Love ℓ_1-regularization.

Andrew Blandino

University of California, Davis
Research Training Group Seminar

June 1st, 2018

Presentation Outline & Goals

By the end of this talk you will, hopefully, be able to answer these questions:

- What is regularization? Why is it useful?

Presentation Outline & Goals

By the end of this talk you will, hopefully, be able to answer these questions:

- What is regularization? Why is it useful?
- What is the Lasso? What are its pros and cons?

Presentation Outline & Goals

By the end of this talk you will, hopefully, be able to answer these questions:

- What is regularization? Why is it useful?
- What is the Lasso? What are its pros and cons?
- How do I use the Lasso for my model?

Presentation Outline & Goals

By the end of this talk you will, hopefully, be able to answer these questions:

- What is regularization? Why is it useful?
- What is the Lasso? What are its pros and cons?
- How do I use the Lasso for my model?
- Which Lasso is right for me and/or my dataset?

Introduction to Regression Model

Recall the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad i = 1, ..., n. \tag{1}$$

where

- ϵ_i: unobservable random errors (typically assume ϵ_i are i.i.d.).

Introduction to Regression Model

Recall the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad i = 1, ..., n. \tag{1}$$

where

- ϵ_i: unobservable random errors (typically assume ϵ_i are i.i.d.).
- Y_i: response for the ith subject.

Introduction to Regression Model

Recall the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad i = 1, ..., n. \tag{1}$$

where

- ϵ_i: unobservable random errors (typically assume ϵ_i are i.i.d.).
- Y_i: response for the ith subject.
- $X_{i1}, ..., X_{ip}$: covariates from the ith subject.

Introduction to Regression Model

Recall the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad i = 1, ..., n. \tag{1}$$

where

- ϵ_i: unobservable random errors (typically assume ϵ_i are i.i.d.).
- Y_i: response for the ith subject.
- $X_{i1}, ..., X_{ip}$: covariates from the ith subject.
- $\beta_1, ..., \beta_p$: coefficients relating covariates to the response, with intercept β_0.

Least Squares Regression

One popular method for fitting this model is using the Least Squares estimator. The Least Squares estimator, $\hat{\beta}$, minimizes objective function

$$\begin{aligned} Q(\boldsymbol{b}) &= Q(b_0, ..., b_p) \\ &= \textstyle\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{i1} - ... - b_p X_{ip}\right)^2 . \end{aligned} \tag{2}$$

Least Squares Regression

One popular method for fitting this model is using the Least Squares estimator. The Least Squares estimator, $\hat{\beta}$, minimizes objective function

$$\begin{aligned} Q(\boldsymbol{b}) &= Q(b_0, ..., b_p) \\ &= \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{i1} - ... - b_p X_{ip}\right)^2 . \end{aligned} \tag{2}$$

Hence, the estimator is defined by:

$$\hat{\beta} = \arg\min_{\boldsymbol{b}} Q(\boldsymbol{b}). \tag{3}$$

Least Squares Regression: Pros

Least-Squares regression is a very popular method for several reasons:

- Tradition: discovered in early 1800's: Gauss, Legendre.

Least Squares Regression: Pros

Least-Squares regression is a very popular method for several reasons:

- Tradition: discovered in early 1800's: Gauss, Legendre.
- Simplicity: solution has explicit form.

Least Squares Regression: Pros

Least-Squares regression is a very popular method for several reasons:

- Tradition: discovered in early 1800's: Gauss, Legendre.
- Simplicity: solution has explicit form.
- Theoretical properties: unbiased, consistent, central limit theorem.

Least Squares Regression: Pros

Least-Squares regression is a very popular method for several reasons:

- Tradition: discovered in early 1800's: Gauss, Legendre.
- Simplicity: solution has explicit form.
- Theoretical properties: unbiased, consistent, central limit theorem.
- Statistical Inference: with normality assumption of residuals, can perform hypothesis tests, construct confidence / prediction intervals, etc.

Least Squares Regression: Cons

There are downsides to using Ordinary Least Squares:

- Too-generous (Lack of Sparsity): estimated coefficients are (practically) never zero.

Least Squares Regression: Cons

There are downsides to using Ordinary Least Squares:

- Too-generous (Lack of Sparsity): estimated coefficients are (practically) never zero.
- Multicollinearity: unstable estimates with correlated independent variables.

Least Squares Regression: Cons

There are downsides to using Ordinary Least Squares:

- Too-generous (Lack of Sparsity): estimated coefficients are (practically) never zero.
- Multicollinearity: unstable estimates with correlated independent variables.
- Overfitting: when $p \approx n$, fits the data 'too-well' i.e. lack of generalization.

Least Squares Regression: Cons

There are downsides to using Ordinary Least Squares:

- Too-generous (Lack of Sparsity): estimated coefficients are (practically) never zero.
- Multicollinearity: unstable estimates with correlated independent variables.
- Overfitting: when $p \approx n$, fits the data 'too-well' i.e. lack of generalization.
- High-dimensional: no unique solution when $p > n$.

Least Squares Regression: Cons

There are downsides to using Ordinary Least Squares:

- Too-generous (Lack of Sparsity): estimated coefficients are (practically) never zero.
- Multicollinearity: unstable estimates with correlated independent variables.
- Overfitting: when $p \approx n$, fits the data 'too-well' i.e. lack of generalization.
- High-dimensional: no unique solution when $p > n$.

These problems motivated researchers to develop alternative methods to address these issues.

Least Squares Regression: Cons

There are downsides to using Ordinary Least Squares:

- Too-generous (Lack of Sparsity): estimated coefficients are (practically) never zero.
- Multicollinearity: unstable estimates with correlated independent variables.
- Overfitting: when $p \approx n$, fits the data 'too-well' i.e. lack of generalization.
- High-dimensional: no unique solution when $p > n$.

These problems motivated researchers to develop alternative methods to address these issues. Hence, the use of regularization.

Regularization example: Curve estimation

- Regularization easily understood in curve estimation:

$$Y_i = f(x_i) + \epsilon_i, i = 1, ..., n.$$

Regularization example: Curve estimation

- Regularization easily understood in curve estimation:

$$Y_i = f(x_i) + \epsilon_i, i = 1, ..., n.$$

- Obtain cubic smoothing spline estimate $\hat{f}$ from minimization of

$$\sum_{i=1}^{n} \left(Y_i - \hat{f}(x_i) \right)^2 + \lambda \int \hat{f}''(x)^2 dx,$$

Regularization example: Curve estimation

- Regularization easily understood in curve estimation:

$$Y_i = f(x_i) + \epsilon_i, i = 1, ..., n.$$

- Obtain cubic smoothing spline estimate $\hat{f}$ from minimization of

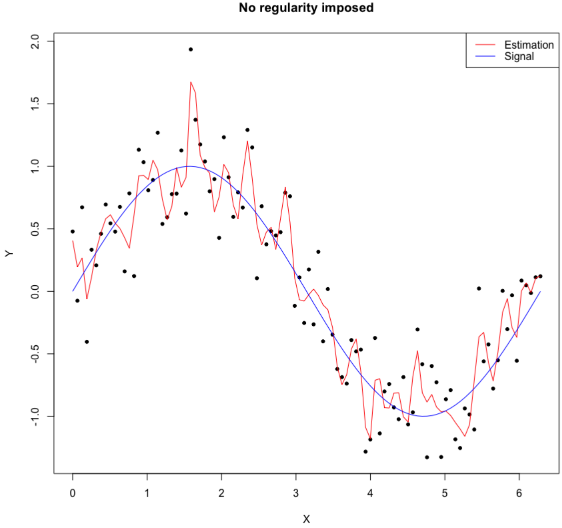
$$\sum_{i=1}^{n} \left(Y_i - \hat{f}(x_i) \right)^2 + \lambda \int \hat{f}''(x)^2 dx,$$

- where

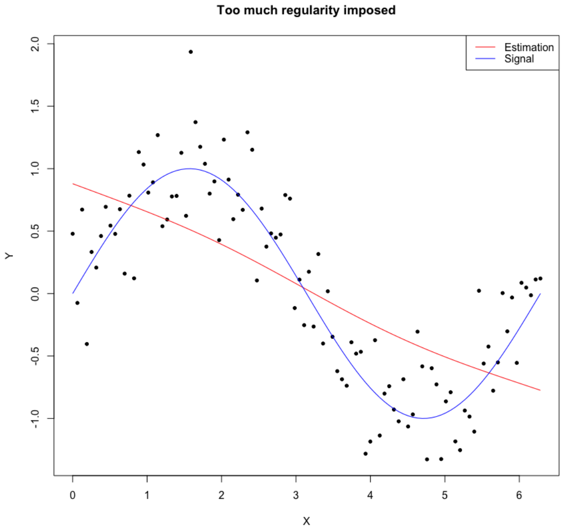
$$\sum_{i=1}^{n} \left(Y_i - \hat{f}(x_i) \right)^2 = \text{Goodness-of-fit}$$

$$\lambda \int \hat{f}''(x)^2 dx = \text{Regularity}$$

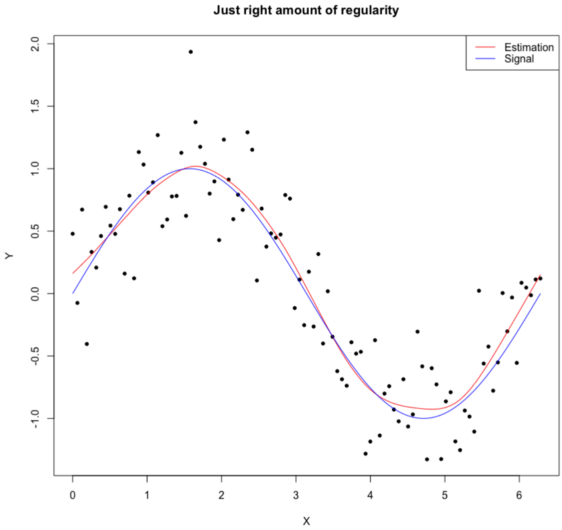
Regularization: No regularity



Regularization: Too much regularity



Regularization: Just right



Ridge Regression

Ridge regression [Hoerl & Kennard (1970)] uses same objective function with constraint:

$$\min_{\boldsymbol{b}} Q(\boldsymbol{b}) \quad \text{subject to} \quad \sum_{i=1}^{p} |b_i|^2 \leq s, \tag{4}$$

where $s \geq 0$ is an additional parameter. Can equivalently write Ridge estimator as

$$\hat{\beta}_{Ridge,\lambda} = \arg\min_{\boldsymbol{b}} \left\{ Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} |b_i|^2 \right\}, \tag{5}$$

where $\lambda \geq 0$ is a tuning parameter.

Ridge Regression

Ridge regression [Hoerl & Kennard (1970)] uses same objective function with constraint:

$$\min_{\boldsymbol{b}} Q(\boldsymbol{b}) \quad \text{subject to} \quad \sum_{i=1}^{p} |b_i|^2 \leq s, \tag{4}$$

where $s \geq 0$ is an additional parameter. Can equivalently write Ridge estimator as

$$\hat{\beta}_{Ridge,\lambda} = \arg\min_{\boldsymbol{b}} \left\{ Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} |b_i|^2 \right\}, \tag{5}$$

where $\lambda \geq 0$ is a tuning parameter. In this form we can see

$$Q(\boldsymbol{b}) = \text{Goodness-of-fit}$$

$$\lambda \sum_{i=1}^{p} |b_i|^2 = \text{Regularity}$$

Ridge Regression

$$\hat{\beta}_{Ridge,\lambda} = \arg\min_{\boldsymbol{b}} \left\{ Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} |b_i|^2 \right\},$$

Because of the added control of the tuning parameter we have the following observations for different values of λ:

- $\lambda = 0$: we obtain the Least Squares solution.

Ridge Regression

$$\hat{\beta}_{Ridge,\lambda} = \arg\min_{\boldsymbol{b}} \left\{ Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} |b_i|^2 \right\},$$

Because of the added control of the tuning parameter we have the following observations for different values of λ:

- $\lambda = 0$: we obtain the Least Squares solution.
- $\lambda = \infty$: we obtain the trivial solution, $\boldsymbol{b} = \mathbf{0}$.

Ridge Regression

$$\hat{\beta}_{Ridge,\lambda} = \arg\min_{\boldsymbol{b}} \left\{ Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} |b_i|^2 \right\},$$

Because of the added control of the tuning parameter we have the following observations for different values of λ:

- $\lambda = 0$: we obtain the Least Squares solution.
- $\lambda = \infty$: we obtain the trivial solution, $\boldsymbol{b} = \mathbf{0}$.
- $0 < \lambda < \infty$: we have shrunk version of the Least Squares solution.

Ridge Regression

$$\hat{\beta}_{Ridge,\lambda} = \arg\min_{\boldsymbol{b}} \left\{ Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} |b_i|^2 \right\},$$

Because of the added control of the tuning parameter we have the following observations for different values of λ:

- $\lambda = 0$: we obtain the Least Squares solution.
- $\lambda = \infty$: we obtain the trivial solution, $\boldsymbol{b} = \mathbf{0}$.
- $0 < \lambda < \infty$: we have shrunk version of the Least Squares solution.

Hence, λ is also called a shrinkage parameter.

Inspection of Constraint

A few comments on the constraint posed:

$$\sum_{i=1}^{p} |b_i|^2 \leq s.$$

Inspection of Constraint

A few comments on the constraint posed:

$$\sum_{i=1}^{p} |b_i|^2 \leq s.$$

- No constraint on intercept b_0.

Inspection of Constraint

A few comments on the constraint posed:

$$\sum_{i=1}^{p} |b_i|^2 \leq s.$$

- No constraint on intercept b_0.
- Scale of covariates X_i will affect solution.

Inspection of Constraint

A few comments on the constraint posed:

$$\sum_{i=1}^{p} |b_i|^2 \leq s.$$

- No constraint on intercept b_0.
- Scale of covariates X_i will affect solution.

Hence, we typically center and standardize the covariates (X_i's), and center the response (Y_i's).

Summary of Ridge regression

Pros:

- High-dimensional: unique solution when $p > n$.

Summary of Ridge regression

Pros:

- High-dimensional: unique solution when $p > n$.
- Multicollinearity: works well in spite of multicollinear data.

Summary of Ridge regression

Pros:

- High-dimensional: unique solution when $p > n$.
- Multicollinearity: works well in spite of multicollinear data.
- Better fit: doesn't overfit when $p \approx n$ (with proper λ).

Summary of Ridge regression

Pros:

- High-dimensional: unique solution when $p > n$.
- Multicollinearity: works well in spite of multicollinear data.
- Better fit: doesn't overfit when $p \approx n$ (with proper λ).
- Closed-form: calculate standard errors, asymptotics, etc.

Summary of Ridge regression

Pros:

- High-dimensional: unique solution when $p > n$.
- Multicollinearity: works well in spite of multicollinear data.
- Better fit: doesn't overfit when $p \approx n$ (with proper λ).
- Closed-form: calculate standard errors, asymptotics, etc.

Cons:

Summary of Ridge regression

Pros:

- High-dimensional: unique solution when $p > n$.
- Multicollinearity: works well in spite of multicollinear data.
- Better fit: doesn't overfit when $p \approx n$ (with proper λ).
- Closed-form: calculate standard errors, asymptotics, etc.

Cons:

- Biased: for $\lambda > 0$.

Summary of Ridge regression

Pros:

- High-dimensional: unique solution when $p > n$.
- Multicollinearity: works well in spite of multicollinear data.
- Better fit: doesn't overfit when $p \approx n$ (with proper λ).
- Closed-form: calculate standard errors, asymptotics, etc.

Cons:

- Biased: for $\lambda > 0$.
- Too-generous (lack of sparsity): like OLS, estimated coefficients are (practically) never zero.

Introduction to Lasso

Tibshirani (1996) introduced the Least Absolute Shrinkage and Selection Operator

$$\min_{\boldsymbol{b}} Q(\boldsymbol{b}) \text{ subject to } \sum_{i=1}^{p} |b_i| \leq s \tag{6}$$

for some $s > 0$. Or, equivalently,

$$\hat{\beta}_{Lasso,\lambda} = \arg\min_{\boldsymbol{b}} \left\{ Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} |b_i| \right\},$$

for tuning parameter $\lambda > 0$.

Difference between Lasso and Ridge

Notice that the only difference between the Lasso and Ridge is the 'loss' used for the penalty, i.e. both have constraint of the form

$$\sum_{i=1}^{p} l(b_i) \leq s.$$

- Lasso: $l(b_i) = |b_i|$ (ℓ_1-penalty)
- Ridge: $l(b_i) = |b_i|^2$ (ℓ_2-penalty)

This seemingly minor detail has major ramifications towards the utility and popularity of the Lasso.

Advantages of Lasso

- Variable Selection (Sparsity): can estimate regression coefficients as 0 i.e. principle of parsimony.

Advantages of Lasso

- Variable Selection (Sparsity): can estimate regression coefficients as 0 i.e. principle of parsimony.
- Fast: can perform variable selection much faster than best-subsets for large number of predictors (LARS algorithm).

Advantages of Lasso

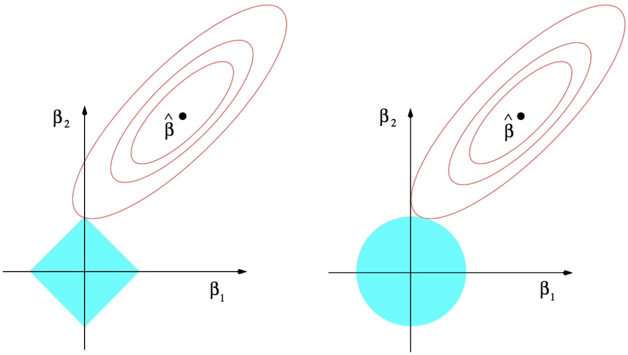
- Variable Selection (Sparsity): can estimate regression coefficients as 0 i.e. principle of parsimony.
- Fast: can perform variable selection much faster than best-subsets for large number of predictors (LARS algorithm).
- Better fit: doesn't overfit when $p \approx n$ (with proper λ).

Advantages of Lasso

- Variable Selection (Sparsity): can estimate regression coefficients as 0 i.e. principle of parsimony.
- Fast: can perform variable selection much faster than best-subsets for large number of predictors (LARS algorithm).
- Better fit: doesn't overfit when $p \approx n$ (with proper λ).
- Valid in High-Dimensions: works for $p > n$.

Comparison between Lasso and Ridge

(Graphic from Tibshirani)



Lasso $\sum_j |\beta_j| \leq s$ Ridge $\sum_j \beta_j^2 \leq s$

Bayesian Interpretation

- The Lasso can be viewed as a Bayes estimate.

Bayesian Interpretation

- The Lasso can be viewed as a Bayes estimate.
- Put Laplace (Double exponential) priors on each β_j

$$\beta_j \overset{i.i.d.}{\sim} f(\beta_j) = \frac{\lambda}{2} \exp\left(-\lambda \left|\beta_j\right|\right)$$

Bayesian Interpretation

- The Lasso can be viewed as a Bayes estimate.
- Put Laplace (Double exponential) priors on each β_j

$$\beta_j \overset{i.i.d.}{\sim} f(\beta_j) = \frac{\lambda}{2} \exp\left(-\lambda \left|\beta_j\right|\right)$$

- Then, for errors $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, we get

$$\hat{\beta}_{Lasso} = \arg\max_{\beta} p(\beta | Y, \sigma^2, \lambda).$$

Bayesian Interpretation

- The Lasso can be viewed as a Bayes estimate.
- Put Laplace (Double exponential) priors on each β_j

$$\beta_j \overset{i.i.d.}{\sim} f(\beta_j) = \frac{\lambda}{2} \exp\left(-\lambda \left|\beta_j\right|\right)$$

- Then, for errors $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, we get

$$\hat{\beta}_{Lasso} = \arg\max_{\beta} p(\beta | Y, \sigma^2, \lambda).$$

- Similarly, for Ridge regression

$$\hat{\beta}_{Ridge} = \arg\max_{\beta} p(\beta | Y, \sigma^2, \lambda),$$

Bayesian Interpretation

- The Lasso can be viewed as a Bayes estimate.
- Put Laplace (Double exponential) priors on each β_j

$$\beta_j \overset{i.i.d.}{\sim} f(\beta_j) = \frac{\lambda}{2} \exp\left(-\lambda \left|\beta_j\right|\right)$$

- Then, for errors $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, we get

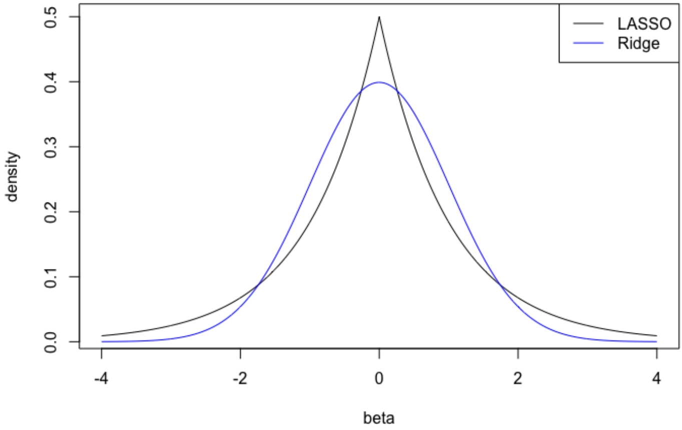
$$\hat{\beta}_{Lasso} = \arg\max_{\beta} p(\beta | Y, \sigma^2, \lambda).$$

- Similarly, for Ridge regression

$$\hat{\beta}_{Ridge} = \arg\max_{\beta} p(\beta | Y, \sigma^2, \lambda),$$

when $\beta_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \lambda)$.

Bayesian Interpretation



Disadvantages of Lasso

- Biased: unless $\lambda = 0$.

Disadvantages of Lasso

- Biased: unless $\lambda = 0$.
- Numerical Solution: due to ℓ_1 penalty, closed-form solution not possible.

Disadvantages of Lasso

- Biased: unless $\lambda = 0$.
- Numerical Solution: due to ℓ_1 penalty, closed-form solution not possible.
- Statistical Inference: harder to perform hypothesis tests, confidence intervals, etc.

Disadvantages of Lasso

- Biased: unless $\lambda = 0$.
- Numerical Solution: due to ℓ_1 penalty, closed-form solution not possible.
- Statistical Inference: harder to perform hypothesis tests, confidence intervals, etc.
- Multicollinearity: will select correlated predictors 'randomly'.

Choosing λ

- Depends on goals of modeling (prediction error, inference, etc.).

Choosing λ

- Depends on goals of modeling (prediction error, inference, etc.).
- Cross-validation (CV): most popular for estimating prediction error.

Choosing λ

- Depends on goals of modeling (prediction error, inference, etc.).
- Cross-validation (CV): most popular for estimating prediction error.
- Generalized Cross-validation (GCV): extension on Cross-validation.

Choosing λ

- Depends on goals of modeling (prediction error, inference, etc.).
- Cross-validation (CV): most popular for estimating prediction error.
- Generalized Cross-validation (GCV): extension on Cross-validation.
- Information Criteria: AIC, BIC, MDL etc.

Prostate Data

Prostate Data (Stamey et. al): interested in associating level of prostate-specific antigen (`lpsa`) with following clinical measures:

- `lcavol`: log cancer volume.
- `lweight`: log prostate weight.
- `age`: patient's age.
- `lbph`: log of amount of benign prostate hyperplasia.
- `svi`: seminal vesicle invasion.
- `lcp`: log of capsular penetration.
- `gleason`: Gleason score.
- `pgg45`: percent of Gleason scores 4 or 5.

97 patients, then randomly split into training group (67) and testing group (30).

Prostate Data: Comparison

Term	LS	Best Subset	Ridge	Lasso
Intercept	2.45	2.45	2.45	2.45
lcavol	0.716	0.78	0.604	0.562
lweight	0.293	0.352	0.286	0.189
age	-0.143	0	-0.108	0
lbph	0.212	0	0.201	0.003
svi	0.31	0	0.283	0.096
lcp	-0.289	0	-0.154	0
gleason	-0.021	0	0.014	0
pgg45	0.277	0	0.203	0
Test Error	0.549	0.548	0.517	0.453

Evolutions of Lasso: Adaptive Lasso

- Adaptive Lasso (Zou, 2006) modified ℓ_1-penalty:

$$Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} \frac{|b_i|}{\left(\left|\hat{\beta}_i^*\right|\right)^{\gamma}},$$

where $\hat{\beta}_i^*$ are initial estimates of β_i (say, from Ridge).

Evolutions of Lasso: Adaptive Lasso

- Adaptive Lasso (Zou, 2006) modified ℓ_1-penalty:

$$Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} \frac{|b_i|}{\left(\left|\hat{\beta}_i^*\right|\right)^{\gamma}},$$

where $\hat{\beta}_i^*$ are initial estimates of β_i (say, from Ridge).

  - Asymptotic Normality,

Evolutions of Lasso: Adaptive Lasso

- Adaptive Lasso (Zou, 2006) modified ℓ_1-penalty:

$$Q(\boldsymbol{b}) + \lambda \sum_{i=1}^{p} \frac{|b_i|}{\left(\left|\hat{\beta}_i^*\right|\right)^{\gamma}},$$

where $\hat{\beta}_i^*$ are initial estimates of β_i (say, from Ridge).

  - Asymptotic Normality,
  - Selection Consistency.

Evolutions of Lasso: Fused Lasso

- Fused Lasso: for data with an inherent-ordering, Tibshirani et. al (2005) proposed the following modification:

$$\min_{\boldsymbol{b}} Q(\boldsymbol{b}) \quad \text{subject to} \begin{cases} \sum_{i=1}^{p} |b_i| \leq s_1 \\ \sum_{i=2}^{p} |b_i - b_{i-1}| \leq s_2 \end{cases}$$

Evolutions of Lasso: Fused Lasso

- Fused Lasso: for data with an inherent-ordering, Tibshirani et. al (2005) proposed the following modification:

$$\min_{\boldsymbol{b}} Q(\boldsymbol{b}) \quad \text{subject to} \begin{cases} \sum_{i=1}^{p} |b_i| \leq s_1 \\ \sum_{i=2}^{p} |b_i - b_{i-1}| \leq s_2 \end{cases}$$

- Sparsity in coefficients *and* their differences.

Evolutions of Lasso: Fused Lasso

- Fused Lasso: for data with an inherent-ordering, Tibshirani et. al (2005) proposed the following modification:

$$\min_{\boldsymbol{b}} Q(\boldsymbol{b}) \quad \text{subject to} \begin{cases} \sum_{i=1}^{p} |b_i| \leq s_1 \\ \sum_{i=2}^{p} |b_i - b_{i-1}| \leq s_2 \end{cases}$$

- Sparsity in coefficients *and* their differences.
- E.g. Spectrometry data, graphical models, etc.

Evolutions of Lasso: Fused Lasso

- Fused Lasso: for data with an inherent-ordering, Tibshirani et. al (2005) proposed the following modification:

$$\min_{\boldsymbol{b}} Q(\boldsymbol{b}) \quad \text{subject to} \begin{cases} \sum_{i=1}^{p} |b_i| \leq s_1 \\ \sum_{i=2}^{p} |b_i - b_{i-1}| \leq s_2 \end{cases}$$

- Sparsity in coefficients *and* their differences.
- E.g. Spectrometry data, graphical models, etc.
- Can outperform Lasso with ordered data.

How Do I Lasso my dataset?

- (R) `glmnet`: fits (general) linear models (including other regression models: logistic, multinomial, etc.) with Elastic-Net (mixture of Ridge and Lasso).
- (R) `monomvn`: Bayesian Lasso.
- (SAS) PROC GLMSELECT: by specifying the model selection method to use Lasso (SELECTION=Lasso).
- (STATA) LassoPACK: fits Lasso, Ridge, A-Lasso, and also does K-fold cross-validation.

Bibliography & References I

A. Ahrens, C. B. Hansen, and M. E. Schaffer.
LASSOPACK: Stata module for lasso, square-root lasso, elastic net, ridge, adaptive lasso estimation and cross-validation.
Statistical Software Components, Boston College Department of Economics, Feb. 2018.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani.
LEAST ANGLE REGRESSION.
The Annals of Statistics, 32(2):407–499, 2004.

T. Hastie, R. Tibshirani, and J. Friedman.
The Elements of Statistical Learning.
2009.

Bibliography & References II

A. E. Hoerl and R. W. Kennard.
Ridge Regression: Biased Estimation for Nonorthogonal Problems.
Technometrics, 12(1):55, feb 1970.

T. Park and G. Casella.
The bayesian lasso.
Journal of the American Statistical Association, 103(482):681–686, 2008.

R. Tibshirani.
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

Bibliography & References III

R. Tibshirani.
The Lasso: A brief Review and a new significance test, 2014 (accessed June 1, 2018).
http://statweb.stanford.edu/~tibs/ftp/ubctalk.pdf.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight.
Sparsity and smoothness via the fused lasso.
J. R. Statist. Soc. B, 67(1):91–108, 2005.

H. Zou.
The adaptive lasso and its oracle properties.
Journal of the American Statistical Association, 101(476):1418–1429, 2006.