The LASSO

Least Absolute Shrinkage and Selection Operator

Zubia Mansoor

November 27, 2019

Overview

Motivation for the LASSO

Ordinary Least Squares Estimates

The OLS estimates are obtained by minimizing the residual squared error.

But why are we not satisfied with them?

Ordinary Least Squares Estimates

The OLS estimates are obtained by minimizing the residual squared error.

But why are we not satisfied with them?

- **Prediction Accuracy:**
  - OLS estimates often have low bias but large variance
  - This can be improved by shrinking or setting some coefficients to 0

Ordinary Least Squares Estimates

The OLS estimates are obtained by minimizing the residual squared error.

But why are we not satisfied with them?

- **Prediction Accuracy:**
  - OLS estimates often have low bias but large variance
  - This can be improved by shrinking or setting some coefficients to 0
- **Interpretability**
  - Large number of predictors lead to less interpretable models
  - We would rather determine a smaller subset of features exhibiting the strongest effects

Subset Selection and Ridge Regression

Subset Selection

- Retains only a subset of the variables
- Discards the rest from the model
  - **More Interpretability:** Fewer predictors in the model
  - **Low Prediction Accuracy:** Extremely variable since it is a discrete process

Subset Selection and Ridge Regression

Subset Selection

- Retains only a subset of the variables
- Discards the rest from the model
  - **More Interpretability:** Fewer predictors in the model
  - **Low Prediction Accuracy:** Extremely variable since it is a discrete process

Ridge Regression

- Shrinks the regression coefficients by imposing a penalty on their size
- Ridge coefficients minimize a penalized residual sum of squares
  - **More Stability:** Continuous process and doesn't suffer as much from high variability
  - **Less Interpretability:** Does not set any coefficients to 0

The LASSO

- Regression analysis method that is able to achieve both of these goals:
  - **Variable Selection** $\rightarrow$ More Interpretable Models
  - **Regularization** $\rightarrow$ Enhanced Prediction Accuracy
- Popularized in 1996 by Robert Tibshirani
- Shrinks some coefficients and sets others to 0
- Falls somewhere between ridge regression and subset selection : enjoys some of the properties of each

Model

Model Assumptions

- Suppose we have data $(x_i, y_i), i = 1, 2, \ldots, N$, where $x_i = (x_{i_1}, \ldots, x_{i_p})^T$ are the predictor variables and y_i are the responses
- Observations are independent or $y_i's$ are conditionally independent given the $x_{ij}'s$
- $x_{ij}'s$ are standardized so that

$$\sum_{i=1}^{N} \frac{x_{ij}}{N} = 0 \qquad \text{and} \qquad \sum_{i=1}^{N} \frac{x_{ij}^2}{N} = 1$$

- The design matrix need not be of full rank

Model: Constraint Form

- The lasso estimate is defined by:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \qquad subject \qquad to \qquad \sum_j |\beta_j| \leq t \qquad (1)$$

- Quadratic programming problem with linear inequality constraints
- **The tuning parameter t** controls the amount of shrinkage that is applied to the estimates
- For all t, the solution for : $\beta_0 = \hat{\beta}_0 = \bar{y}$

Model: Penalized Form

- The lasso problem can be framed in the equivalent **Lagrangian form:**

$$\hat{\beta}_{lasso} = \arg\min_{\beta}\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|\} \tag{2}$$

- **The tuning parameter λ** controls the amount of penalty:
  - $\lambda = 0 \rightarrow$ no parameters are eliminated, hence, $\hat{\beta}_{lasso} = \hat{\beta}_{OLS}$
  - As λ increases $\rightarrow$ more and more coefficients are set to zero and eliminated
  - $\lambda = \infty \rightarrow$ all coefficients are eliminated.

- λ has a one-to-one correspondence with t

Sparsity of LASSO

Orthonormal Design Case

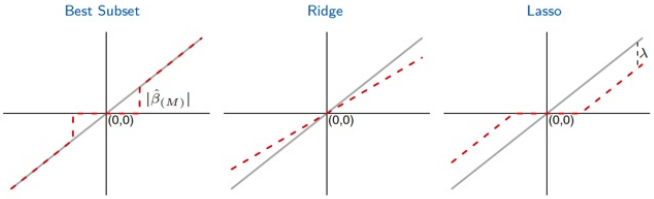
- Let X be the $n \times p$ design matrix with ij^{th} entry x_{ij}
- Suppose $X^T X = I$, the identity matrix
- The solutions to equation (1) are shown to be:

$$\hat{\beta}_j = \textit{sign}(\hat{\beta}_j^o)(|\hat{\beta}_j^o| - \lambda)_+ \tag{3}$$

where,

- $\hat{\beta}_j^{o\prime} s$ are the OLS estimates
- λ is a constant chosen by the corresponding technique
- sign denotes the sign of its argument (± 1)
- x_+ denotes the "positive part" of x

Geometric form of the functions



- Best subset selection of size M drops all variables with coefficients smaller than the M^{th} largest
- Ridge regression scales the coefficient by a constant factor
- LASSO translates by a constant factor, truncating at 0

Constraint regions

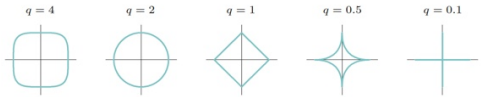
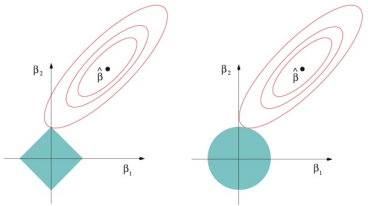


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

- The constraint region for ridge regression is the disk $\beta_1^2 + \beta_2^2 \leq t$
- The constraint region for LASSO is the diamond $|\beta_1| + |\beta_2| \leq t$
- $q = 1$**:** the smallest value of q giving a convex region which is convenient for optimization
- $q > 1$**:** $|\beta_j|^q$ is differentiable at 0, hence, doesn't share the ability of lasso ($q = 1$) for setting coefficients exactly to 0

General Case



- **Residual sum of squares** $\rightarrow$ elliptical contours centered at the full least square estimates
- **LASSO solution** $\rightarrow$ first place where contours touch the diamond; if solution occurs at a corner, it corresponds to a zero solution
- **Ridge solution** $\rightarrow$ No corners for the contours to hit; zero solutions will rarely result
- $\boldsymbol{p} > 2$ $\rightarrow$ the diamond becomes a rhomboid; more opportunities for the estimated parameters to be zero

Illustration

Prostrate Cancer Data

- Interested in the level of prostate-specific antigen (PSA), elevated in men who have prostate cancer
- $y_i = \log(PSA), x_{ij}$ measurements on a man and his prostate
- **Data:** n = 97 men with prostate cancer and p = 8 clinical predictors
- What if we want to derive a linear model using only a few of the 8 predictors to predict the level of PSA?

Model Evaluation

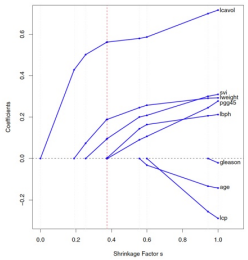


TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

- Shrinkage factor $s = t/\sum_j |\hat{\beta}_j^o|$
- Vertical line is drawn at $s = 0.36$, optimal value chosen by cross-validation
- LASSO profiles hit zero, while those for ridge would not
- Lasso has the lowest test error. Hence, performs better than Subset, Ridge

Comparison between Subset Selection, Ridge, LASSO

Comparison between Subset Selection, Ridge, LASSO

Better Prediction Accuracy

More Interpretable Models

Comparison between Subset Selection, Ridge, LASSO

Better Prediction Accuracy *More Interpretable Models*

- **Small number of large effects:** Subset selection does best here, the lasso not quite as well and ridge regression does quite poorly

Comparison between Subset Selection, Ridge, LASSO

Better Prediction Accuracy *More Interpretable Models*

- **Small number of large effects:** Subset selection does best here, the lasso not quite as well and ridge regression does quite poorly

- **Small to moderate number of moderate-sized effects:** Lasso does best, followed by ridge regression and then subset selection

Comparison between Subset Selection, Ridge, LASSO

Better Prediction Accuracy *More Interpretable Models*

- **Small number of large effects:** Subset selection does best here, the lasso not quite as well and ridge regression does quite poorly
- **Small to moderate number of moderate-sized effects:** Lasso does best, followed by ridge regression and then subset selection
- **Large number of small effects:** Ridge regression does best by a good margin, followed by the lasso and then subset selection

Summary

Summary

- We need better techniques to handle high-dimensional data with better prediction accuracy and more interpretability
- LASSO makes use of the L_1 penalty to set some coefficents exactly to 0 and gives us sparse models
- Some Applications of LASSO include:
  - **Tree-based methods:** Rather than prune a large tree, use the LASSO idea to shrink it
  - **Multivariate Adaptive Regression Splines:** Special Lasso-type algorithm to grow and prune a MARS model

References

- **Regression Shrinkage and Selection via the Lasso**, R. Tibshirani. Journal of the Royal Statistical Society. Series B (Methodological) (1996) (Methodological), Vol. 58, No. 1(1996), pp. 267-288

- **Regression shrinkage and selection via the lasso: a retrospective**, R. Tibshirani. Journal of the Royal Statistical Society Series B, 2011, vol. 73, issue 3, 273-282

- **The Elements of Statistical Learning**, T. Hastie, R. Tibshirani, and J. Friedman. Springer Series in Statistics Springer New York Inc., New York, NY, USA, (2001)

Questions?