Lasso Regression

Cristián Aguilera-Arellano
UMN

February 8, 2021

Introduction

- Least squares estimates often have low bias but large variance
  - Prediction accuracy might improve by shrinking or setting some coefficients to zero

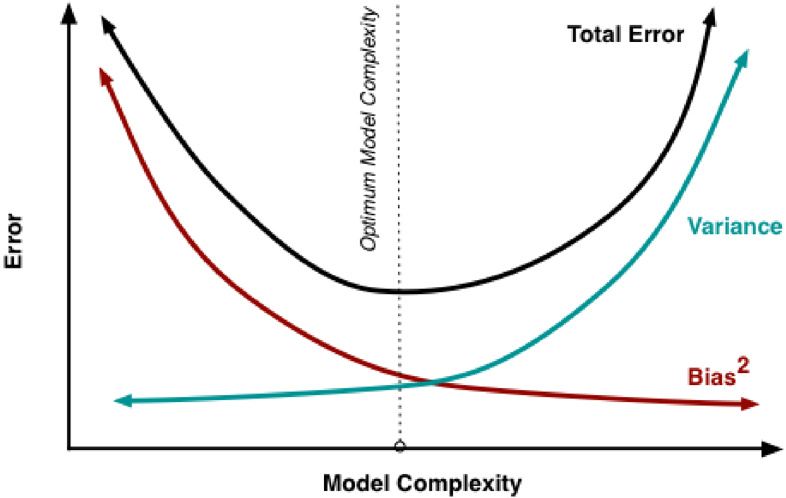
- The mean squared error of an estimator $\tilde{\beta}$

$$MSE(\tilde{\beta}) = E(\tilde{\beta} - \beta)^2$$
$$MSE(\tilde{\beta}) = Var(\tilde{\beta}) + \underbrace{[E(\tilde{\beta}) - \beta]^2}_{Bias}$$

- Gauss-Markov theorem $\longrightarrow$ Least square estimator has the smallest *MSE* of all linear estimators with no bias

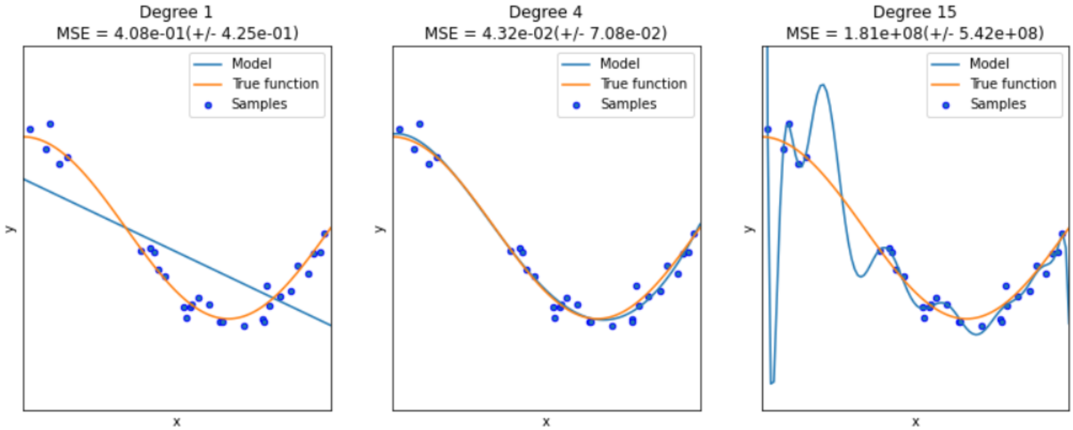
- May exist biased estimators with smaller mean squared error $\rightarrow$ trade a little bias for a larger reduction in variance

Bias-Variance trade-off



Example

- The objective is to create a model that has the best out of sample prediction



Lasso

- Lasso (least absolute shrinkage and selection operator) is a shrinkage method

- The lasso estimate is defined by

$$\hat{\beta}^{lasso} = \text{argmin}_{\beta} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2$$

$$\text{subject to} \qquad \underbrace{\sum_{j=1}^{p} |\beta_j| \le t}_{\text{if it is replaced with } (\beta_j)^2 \text{ then it is called a Ridge regression}}$$

- Making t sufficiently small will cause some of the coefficients to be exactly zero

- We can tune t to minimize the MSE $\rightarrow$ will help to avoid over-fitting

- If we choose $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j^{ls}|$, then the lasso estimates are also the least squares coefficients

When to use Lasso?

- If we have too many variables (p) relative to the number of observations (n)
- If we are willing to increase the bias of the estimates with the objective to reduce the mean squared errors
- If we want a subset of predictors that can produce an interpretable model

Standardize data

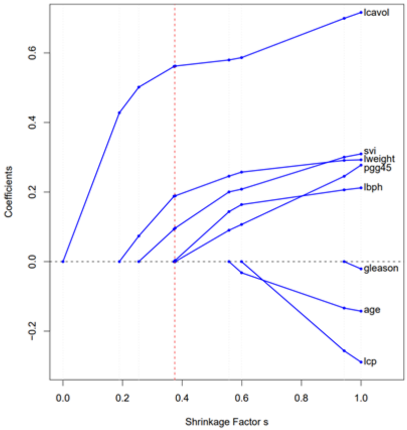
- Since we are penalizing the coefficients of the regression it is important to standardize the predictors

$$\tilde{x} = \frac{x - \bar{x}}{\sigma_x}$$

- This ensures all inputs are treated equally in the regularization process

Example-Prostate cancer

- Objective: Predict the prostate-specific antigen levels
- Predictors: log cancer volume (lcavol), log prostate weight (lweight), age, etc.
- $s = t / \sum_{j=1}^{p} |\hat{\beta}_j|$



Optimal t

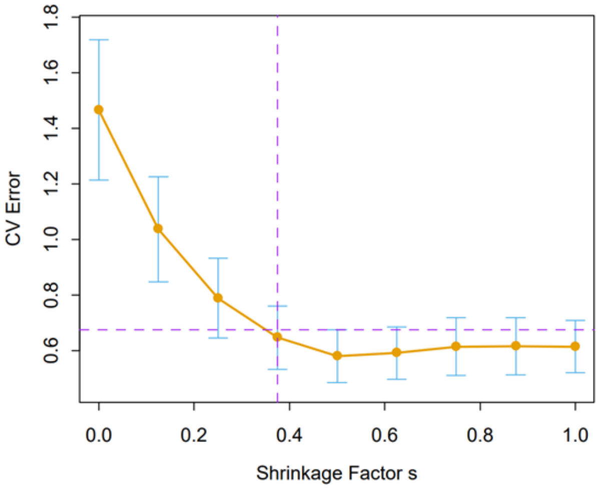
- To determine the optimal $t \longrightarrow$ 10-fold cross-validation
- Randomly select 9 of the 10 folds to train the algorithm and the remaining fold as a test-fold
- After predicting the output in the test-fold, repeat so that each cross-validation fold is used once as a test-fold
- Let $\kappa : \{1, ..., N\} \rightarrow \{1, ..., 10\}$ be a function that indicates the partition to which observation i is allocated
- Denote $\hat{f}^{-k}(x)$ the fitted function, computed with the $k \in \{1, ..., 10\}$ test-fold
- The cross-validation estimate of prediction error is

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

- Select the t^* following the "one-standard error" rule $\rightarrow$ choose the most parsimonious model whose error is no more than one standard error above the error of the best model

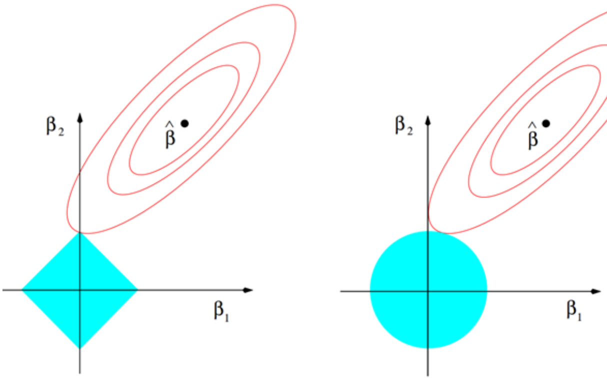
Cross-validation prediction error

- When $s = 1$ the coefficients are the least squares estimates



Ridge regression and Lasso

- The only difference is that the constraint for the ridge regression is: $\sum_{j=1}^{p}(\beta_j)^2 \leq t$
- Blue areas represent the constraints of each problem (lasso (left) and ridge (right))
- The red ellipses are the errors of the least squared error function
- The main difference is that if the solution in lasso hits a corner, one β_j will equal zero



Elastic Net

- Zou and Hastie (2005) introduced the elastic net penalty

$$\sum_{j=1}^{p} \alpha|\beta_j| + (1-\alpha)(\beta_j)^2 \leq t$$

- α determines the mix of the penalties – we can choose α and t by cross-validation
- It shrinks the coefficients of correlated predictors like ridge, and selects variables like lasso

R package

- **glmnet** – package that fits a generalized linear model via penalized maximum likelihood
- The regularization path is computed for the lasso, elastic net or ridge penalty at a grid of values for t
- glmnet algorithm uses cyclical coordinate descent – successively optimizes the objective function over each parameter, and cycles until convergence

Conclusions

- Bias-Variance trade-off
- Tune the parameter t to avoid over-fitting
- Approaches to regularization - Lasso and Ridge regression

References

- Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- https://web.stanford.edu/ hastie/glmnet/glmnet_alpha.html