Research Module in Econometrics and Statistics

A Comparative Study of Shrinkage Methods- LASSO, Ridge & Elastic Net

Manuel Thomas & Nitesh Shisodia
Matriculation Number: 3205373 & 3313402

Under the guidance of:
Prof. Dr. Alois Kneip &
Prof. Dr. Joachim Freyberger

M.Sc. Economics
University of Bonn, Germany, SoSe 2020

Table of Contents

Chapter 1

Introduction

In various applications including in medical sciences, bio-informatics, agricultural sciences, etc., feature selection is a key challenge in model fitting. This is also routinely observed in fields as varied as finance and economics,in estimation problems and machine prediction. Linear models increasingly become unstable when number of parameters(p) approaches sample size(n),and in the presence of highly collinear groups of variables. Fitting over a data set in theses circumstances prior to feature selection can lead to (a)poor accuracy in prediction on new data and (b) uninterpretable models.

Let us consider a linear regression model,with p predictors $x_1...x_p$ and n number of observations. The outcome variable is predicted according to the model(matrix form)-

$$y = \beta_0 + x_1\beta_1... + x_p\beta_p + \epsilon_i$$

A model fitting method which uses linear regression does so by minimizing the the sum of squared residuals, however this suffers from a two pronged failing. Firstly,prediction accuracy: OLS estimates commonly have low bias but a large variance — i.e., the classic blunder of 'Over-fitting'. If we reduce the number of independent variables by reducing their coefficients to zero, we can improve the overall prediction accuracy (Tibshirani, 1996). Secondly, it becomes very difficult to interpret the constructed model in the case of a large number of predictors. Our interest in a task of estimation will be in identifying the most important variables which cumulatively lend the strongest effects.

This can be achieved through various methods,ranging from Lasso(Tibshirani,1996) ,Ridge (Hoerl,A.E.,Kennard, R. W., 1970) to Elastic Net, each with it's key advantages and failings. Lasso attempts to prevent over-fitting by shrinking some coefficients to zero,

according to a model parameter aptly titled 'the shrinkage parameter'. The coefficients are forced to in total not exceed the value of this 't' parameter $t = \sum_{i=1}^{p} \hat{\beta}_i$.

The Chapter 2, starts with the introduction to bias-variance trade-off in the machine learning and explains the theory behind the lasso(with more focus), ridge regression and elastic net. In Chapter 3, we implement Monte Carlo simulations. In the first section, we see how the shrinkage works for lasso and ridge regression. Then, we compare the performances of lasso and ridge under the cases of small and big signals which is followed by performance of all the three methods(including elastic-net) under the correlated explanatory variables. The last section of chapter 3 provides the overall comparison of the methods in varying signals.

Chapter 2

Description of Methods

In this chapter we will start with looking at the ordinary least squares (OLS) estimator and the conditions under which OLS fails. Then, we discuss the theory and assumptions behind LASSO, which is followed by theory behind Ridge and Elastic-net.

2.1 OLS

The usual linear regression in matrix is represented as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where $\beta = (\beta_1, \beta_2, \ldots \beta_p)$ are unknown parameters and ε_{n*1} is the vector of error terms. In order to obtain our prediction estimates, we minimize the residual sum of squares.

Therefore, the objective function is as following:

$$\underset{\beta}{\text{minimize}} \sum (\mathbf{Y} - \mathbf{X}\beta)^2$$

2.1.1 Why OLS fails?

In the traditional setting, the number of observations n is much larger than the number of variables p, that is, $n >> p$. In such scenario, the OLS solution vector is unique, unbiased and consistent.

As the number of predictors p increases, the interpretation of regression models becomes more challenging, especially so for OLS. When the number of variables is much larger than the number of observations, that is $p >> n$, the OLS does not have one unique solution, but infinitely many. These solutions are prone to over-fitting, and while the OLS estimates would in general have a low bias, their large variance adversely affects the prediction accuracy. Now we look into bias and variance in more detail in order to understand how we can use shrinkage methods to improve the prediction accuracy.

2.1.2 Bias vs Variance Trade-off

Bias is defined as the difference between the average prediction of our model and the correct value, which we are trying to predict. A model can have high bias if it does not account for training data enough and oversimplifies the underlying data-generating process.

$$\text{Bias} = E[\hat{\beta}] - \beta$$

Whereas variance is the variability of model prediction for a given data point which shows us the spread of our data. A model with high variance pays a lot of attention to the training data, but does not generalize enough to be able to account for data that has not been used in training; that is, it will perform badly, while predicting the test data.

$$\text{Var}(\hat{\beta}) = E\left[(E[\hat{\beta}] - \hat{\beta})^2\right]$$

For better prediction accuracy, one needs to keep the balance between bias and variance. In the above mentioned case of much larger number of variables then number of observations ($p >> n$), the OLS estimated parameters are unbiased, but have large variance, which results in huge test errors, caused by over-fitting the data.

The expected error of the model can be decomposed in the following manner:

$$E\left[(Y-\hat{Y})^2\right] = [\text{Bias}]^2 + \text{ Variance } + \text{Irreducible Error}$$

The total expected error of the model is represented as the sum of squared bias, variance and the irreducible error. The shrinkage estimation methods aim to substantially reduce the variance of parameters by introducing a small amount of bias, thus improving the overall prediction accuracy of the model.

The two standard techniques that were traditionally used for improving the OLS estimates were subset selection and Ridge regression but they have their drawbacks. In 1996 R. Tibshirani introduced a new method - LASSO. We will first look into the theory and working of LASSO and then have a quick overview of the ridge regression along with elastic net.

2.2 LASSO

As the constraint of minimizing the sum of squared residuals failed, there is need of regularization of the estimation process.

We mostly follow the notations in (Buhlmann and van de Geer, 2011). We consider the usual linear regression setup with univariate response variable $Y \in R$ and p-dimensional variables $X \in R$:

$$Y_i = \sum_{j=1}^{p} X_i^{(j)} \beta_j + \varepsilon_i \quad i = 1, \ldots, n$$

$$\text{where } \varepsilon_i \sim N\left(0, \sigma^2\right)$$

For a quick recap on norms, L1 norm is defined as:

$$\|\beta\|_1 = \Sigma_{j=1}^{p} |\beta_j|$$

L2 norm is defined as:

$$\|\beta\|_2^2 = \Sigma_{j=1}^{p} \beta_j^2$$

The Least Absolute Shrinkage and Selection Operator (LASSO), introduced by (Tibshirani, 1996), is a penalized least squares method that imposes an L1-penalty on the regression coefficients. The LASSO has both the ability to perform shrinkage and automatic variable selection simultaneously due to the nature of the L1-penalty. The LASSO estimator is defined as

$$\hat{\beta}_{\text{LASSO}} = \underset{\beta}{\text{argmin}}(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1) \text{ where } \|\beta\|_1 \leq t$$

Here, $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \sum_{i=1}^{n} (Y_i - (\mathbf{X}\beta)_i)^2$, $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ is l_1 norm of β, $\lambda \geq 0$ is the penalty parameter and t is the user specified parameter which can be thought as budget on the total l_1 norm of the parameter vector and with the help of LASSO we find the best fit within this budget.

If the budget t set is small enough, the LASSO yields sparse solution vectors, having only some coordinates that are non-zero. This is not the same for l_q norms with $q > 1$; for $q < 1$, the solutions here are sparse but the problem is not convex any more, which results in making the minimization very challenging computationally. The value q = 1 is the smallest value that yields a convex problem. Convexity greatly simplifies the computation, as does the sparsity assumption itself. They both together allow for scalable algorithms that can handle problems with even millions of parameters with ease.

2.2.1 Assumptions of Sparsity & Basis Pursuit

LASSO regression does regularization by completely diminishing the importance given to some features (making the weight zero), whereas Ridge regression achieves regularization by reducing the importance given to some of the features and not by nullifying the effects of the features. Due to the same reason, one can say that LASSO regression causes sparsity while Ridge regression does not. Let's see how this actually happens.

In case of $p >> N$ and the true model is not sparse, then the number of observations is too small for accurate estimation of the parameters. But if the true model is sparse, so that only $k < N$ parameters are actually non-zero in the true model, then we can estimate the parameters effectively, using the LASSO and other shrinkage estimation methods.

Basis Pursuit Linear Program uses l_1- norm to solve the equation: $\min_{\beta} ||\beta||_1$ subject to $Y = X\beta$.

Using a simplified example (Gauraha, 2018), we can show how under sparsity assumption under-determined linear equation can be solved. Suppose we have the following equation:

$$\begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Now, assume that the β has a sparse solution. In order to find the solution, we set certain components of the subset of β to zero, such that the above equation holds. Some solutions are:

$$[\beta_1 = 1, \beta_2 = 0, \beta_3 = 0], \|\beta\|_0 = 1, \|\beta\|_1 = 1 \, [\beta_1 = 0, \beta_2 = 2, \beta_3 = 2], \|\beta\|_0 = 2, \|\beta\|_1 = 4$$

We can see that the solution with the least number of non-zero elements is $[\beta_1 = 1, \beta_2 = 0, \beta_3 = 0]$, since $\|\beta\|_0 = 1 < \|\beta\|_0 = 2$. Also, the solution of Basis Pursuit Linear Program is the same for this particular equation; meaning $[\beta_1 = 1, \beta_2 = 0, \beta_3 = 0]$, since $\|\beta\|_1 = 1 < \|\beta\|_1 = 4$.

2.2.2 Orthonormal Design Case

The Lagrange function corresponding to the constrained regression optimization for LASSO is given below:

$$\hat{\beta(\lambda)} = \arg\min_{\beta \in \mathbb{R}^{\text{l}}} \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Since, the objective function above is not differentiable, the LASSO has generally no closed form solution. However, in the case of single-variable model and orthonormal design matrix, a solution is derivable. In the paragraphs that follow, we will look at how to derive a LASSO solution for the single-variable case.

Consider an orthonormal design where p = n and the design matrix $\mathbf{X}$, satisfies $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_{pp} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$. The LASSO estimator then is:

$$\hat{\beta}_j \left(\lambda_1\right) = \text{sign}\left(\hat{\beta}_j\right) \left(\left|\hat{\beta}_j\right| - \frac{1}{2}\lambda_1\right)_+$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{Y}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$. This expression for the LASSO regression estimator can be obtained in the following manner.

We have the LASSO regression loss criterion:

$$\begin{aligned}\min_{\boldsymbol{\beta}} \|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1 &= \min_{\boldsymbol{\beta}} \mathbf{Y}^\top\mathbf{Y}-\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y}+\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}+\lambda_1\sum_{j=1}^{p}|\beta_j| \\ &= \min_{\boldsymbol{\beta}} -\hat{\boldsymbol{\beta}}^\top\boldsymbol{\beta}-\boldsymbol{\beta}^\top\hat{\boldsymbol{\beta}}+\boldsymbol{\beta}^\top\boldsymbol{\beta}+\lambda_1\sum_{j=1}^{p}|\beta_j| \\ &= \min_{\beta_1,\ldots,\beta_p}\sum_{j=1}^{p}\left(-2\hat{\beta}_j^{\text{ot.}}\ \beta_j+\beta_j^2+\lambda_1|\beta_j|\right) \\ &= \sum_{j=1}^{p}\left(\min_{\beta_j} -2\hat{\beta}_j\beta_j+\beta_j^2+\lambda_1|\beta_j|\right)\end{aligned}$$

We can solve the minimization problem with respect to regression coefficient. This gives:

$$\min_{\beta_j} -2\hat{\beta}_j\beta_j+\beta_j^2+\lambda_1|\beta_j| = \begin{cases} \min_{\beta_j} -2\hat{\beta}_j\beta_j+\beta_j^2+\lambda_1\beta_j & \text{if } \beta_j>0 \\ \min_{\beta_j} -2\beta_j\beta_j+\beta_j^2-\lambda_1\beta_j & \text{if } \beta_j<0 \end{cases}$$

The minimization within the sum over the covariates is with respect to each element of the regression parameter separately. Optimization with respect to the j_th gives.

$$\hat{\beta}_j(\lambda_1) = \begin{cases} \hat{\beta}_j-\frac{1}{2}\lambda_1 & \text{if } \hat{\beta}_j(\lambda_1)>0 \\ \hat{\beta}_j+\frac{1}{2}\lambda_1 & \text{if } \hat{\beta}_j(\lambda_1)<0 \\ 0 & \text{otherwise} \end{cases}$$

After putting these two equations together we arrive at the form of the LASSO regression estimator above. The function is also referred to as the soft-threshold function(soft-thresholding refers to the shrinking of coefficients both positive and negative towards zero).

2.3 Ridge Regression

The ad-hoc fix of Hoerl and Kennard (1970) to super-collinearity of the design matrix (and, consequently the singularity of the matrix $\mathbf{X}^\top\mathbf{X}$) has been motivated post-hoc. The ridge estimator minimizes the ridge loss function, which is defined as:

$$\hat{\beta}_{\text{Ridge}} = \underset{\beta}{\text{argmin}}(\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2)$$

The loss function is the traditional sum-of-squares augmented with a penalty. The particular form of the penalty, $\lambda\|\beta\|_2^2$, is referred to as the Ridge penalty and λ as the penalty parameter.

For $\lambda = 0$, minimization of the Ridge loss function yields the least square estimator (if it exists). For any $\lambda > 0$, the ridge penalty contributes to the loss function, affecting its minimum and its location. The $\hat{\beta}_{\text{Ridge}}$ minimizes the sum of-squares and the penalty. The effect of the penalty in this balancing act is to shrink the regression coefficients towards zero, its minimum. In particular, the larger λ, the larger the contribution of the penalty to the loss function, the stronger the tendency to shrink non-zero regression coefficients to zero (and decrease the contribution of the penalty to the loss function). This motivates the name 'penalty' as non-zero elements of increase (or penalize) the loss function.

2.4 Elastic Net

The LASSO can only choose at most n variables that's why it is not the ideal method in high-dimension setting. Empirically, the LASSO some times does not perform well with highly correlated variables. Whereas the ridge regression is unable to produce a parsimonious model as it always includes all the variables in the model. By combining a $l-2$ penalty with the $l-1$ penalty, we obtain the elastic net, another penalized method that deals better with such correlated groups, and tends to select the correlated features (or not) together.

The elastic net makes a compromise between the Ridge and the LASSO penalties (Zou and Hastie 2005); it solves the convex program

$$\hat{\beta}_{\text{ENet}} = \text{argmin}_{\beta}(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda(\alpha||\beta\|_1 + (1-\alpha)\|\beta\|_2^2)$$

where $\alpha \in [0,1]$ is a parameter that can be varied. When $\alpha = 1$, it reduces to the ℓ_1 -norm or LASSO penalty, and with $\alpha = 0$, it reduces to the squared ℓ_2 -norm, corresponding to

the Ridge penalty.

Chapter 3

Simulations

Now after looking at theoretical details of both the LASSO and Ridge regression, we now look at their behavior under different circumstances using the different simulation studies. We will start with the effect of increasing penalty parameter on both LASSO and Ridge. Then we will compare the performance of both the methods under different signals. Then we will compare the performance of both the methods along with the elastic net under different signals and in the case of correlated explanatory variables. At last, we will be comparing the performance of all the methods under varying signals.

3.1 Visualization of LASSO and Ridge with Varying Penalties

The purpose of this simulation set-up is to show visually with the help of graphs, how LASSO and Ridge perform, as the penalty parameter(λ) increases. We start with the base case of OLS and then compare the LASSO and Ridge as λ increases.

The set-up of the simulation is as follows:

- Generate a data set, with $\epsilon \sim N(0,1), \quad \beta^{True} = 0.9, \quad X \sim N(1, 0.65), \quad Y = X\beta^{True} + \epsilon$, $p = 1$, and $n = 100000$.
- Calculate the objective function value, for each $\hat{\beta}$ from a sequence of real values from -0.5 to 1.5 for OLS, LASSO and Ridge.

By the details in theoretical section, we expect that the LASSO sets certain parameters to zero, but Ridge parameters in the same scenario only get asymptotically close to zero.

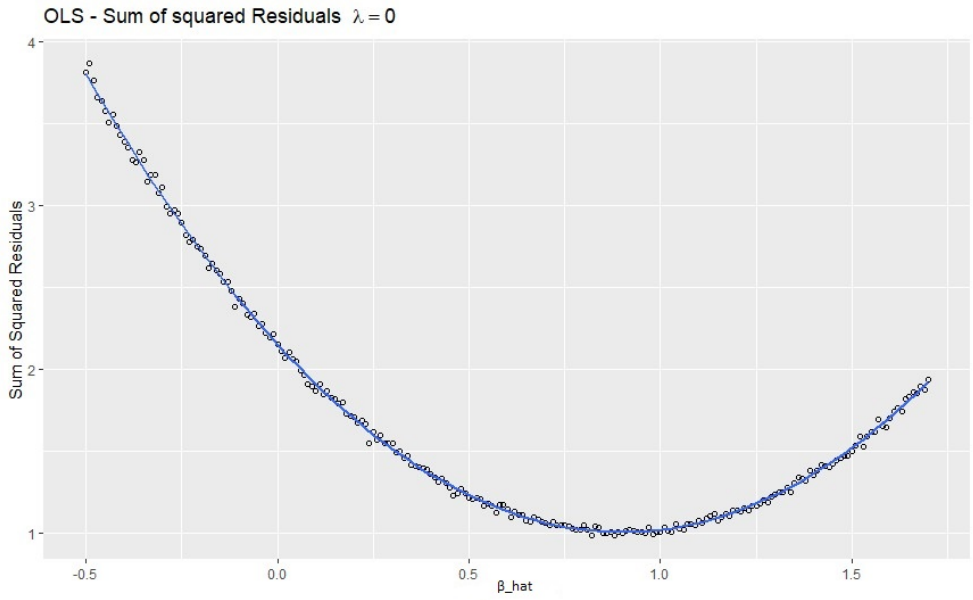

Figure 3.1: Showing RSS values for different $\hat{\beta}$ estimates

In Fig 3.1 the sum of squared residual values are plotted against respective different $\hat{\beta}$. We observe that the minimum RSS is at 0.9, that is, the OLS solution is exactly equal to the $\beta^{True} = 0.9$.

In Fig 3.2 the sum RSS and LASSO penalty is plotted against respective different $\hat{\beta}$ values. The violet curve represents $\lambda = 0$, that is, the the objective function is same as OLS. In that case the objective function is minimum at $\hat{\beta} = 0.9$. But here we observe that as the λ is increased, there is a kink at $\hat{\beta} = 0$, that means if the penalty is substantially high then the parameter value can shrink to 0.

In Fig 3.3 the sum RSS and ridge penalty is plotted against respective different $\hat{\beta}$ values. Again, the violet curve represents $\lambda = 0$, that is, the the objective function is same as OLS.

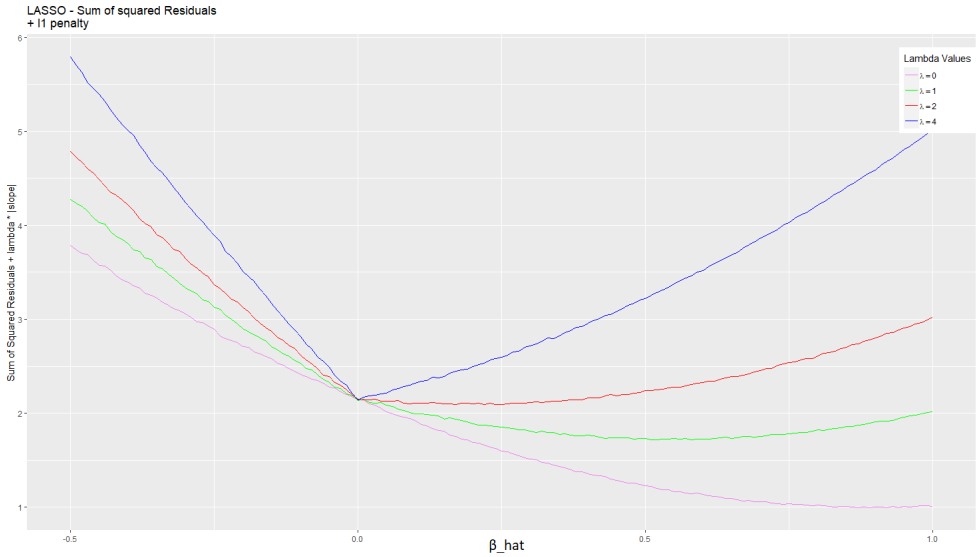

Figure 3.2: Showing the sum of RSS and LASSO penalty for differnt $\hat{\beta}$ and lambda Values

Again, with increase in the penalty parameter, the minimum of the objective function is shrinking. But unlike LASSO, here we observe that as the λ is increased, there is no kink at$\hat{\beta} = 0$, that means even if the penalty is substantially high then also the parameter values can never shrink to 0.

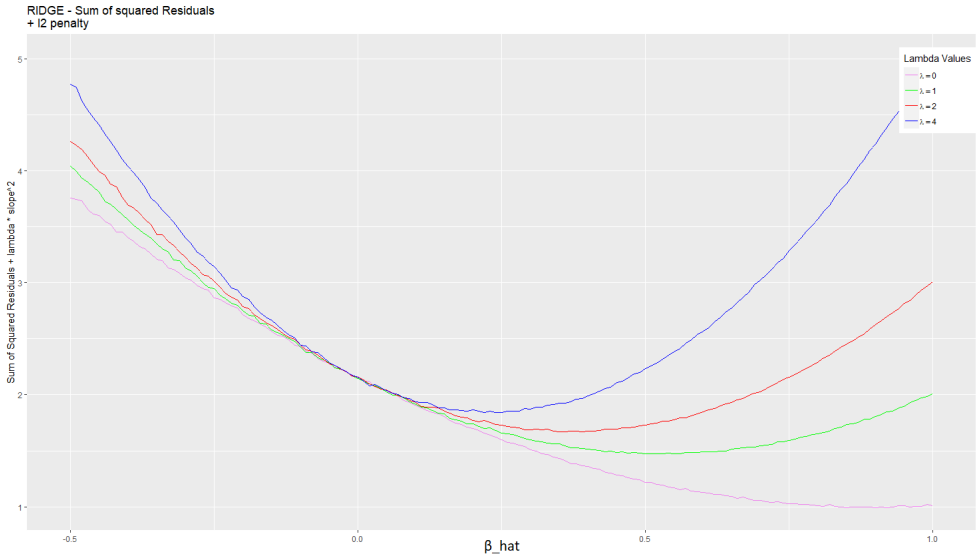

Figure 3.3: Showing the sum of RSS and Ridge penalty for differnt $\hat{\beta}$ and lambda Values

3.2 Comparative Study of LASSO and Ridge in Varying Signals

In the following sections we simulate data to compare the performance of shrinkage methods under different conditions. After seeing which methods work best for certain data set, we try to answer the question why this is the case.

For all the cases below the true model is:

$$Y = X\beta + \epsilon$$

and

$$\epsilon \sim \mathcal{N}_n(0, I),$$

where Y is vector of dependent variable and X is matrix of explanatory variables

3.2.1 Small Signal and Lot of Noise

The β values below indicate the true value of the coefficients in the model. Ans the model specifications for this data generating process is given below:

- $\beta = (\underbrace{1, ..., 1}_{15}, \underbrace{0, ..., 0}_{1485})^T$
- $p = 1500 > n = 500$
- $X_i \overset{\text{iid}}{\sim} N(0, I)$

Here we assume that co-variates are not correlated and there is a small signal, i.e out of 1500 restrictions only 15 are true non-zero parameters. Given the model, we generate the data set and run a regression analysis on the data using LASSO and Ridge. We get the following results shown in 2.

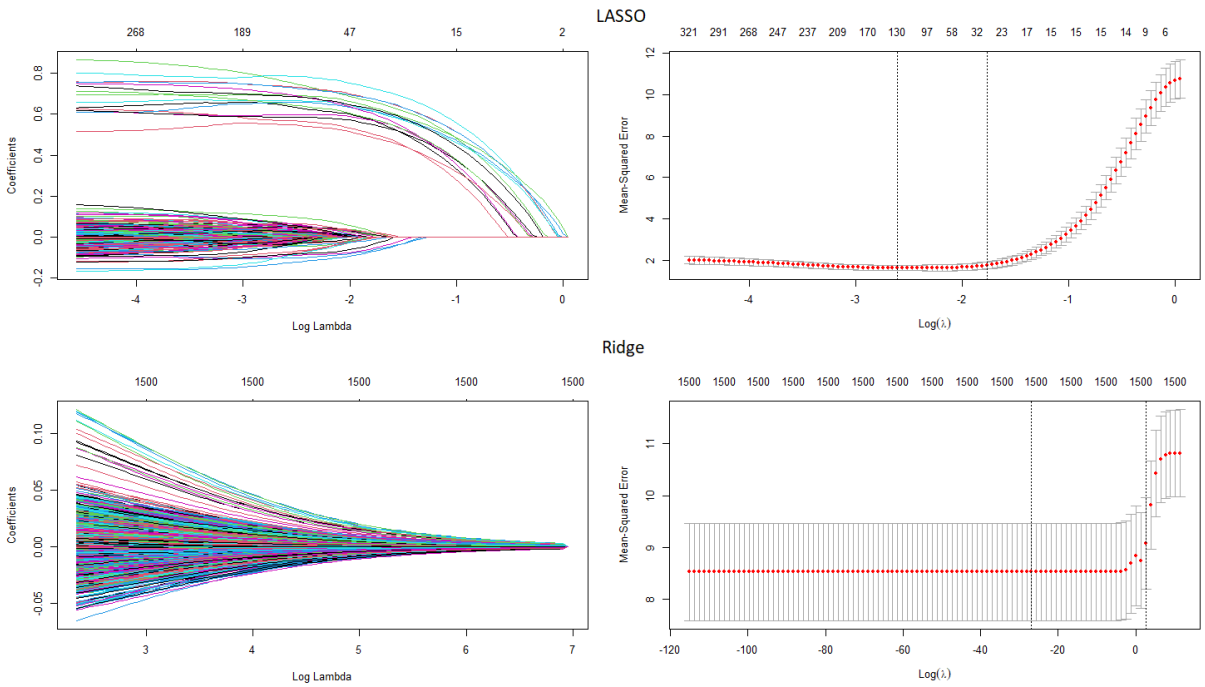

Figure 3.4: Coefficient values against differing log lambda(left side), MSE against varying log(λ)(right side)

From the left part of the graphs we see that LASSO does the best job in variable selection for this data set, considering the true model and the number of true non-zero parameters. The right-hand side graphs illustrate the process of choosing the λ and corresponding model, which minimise Mean Squared Error (MSE), through cross-validation. To see, which shrinkage method is more accurate in this case, we calculate MSE for chosen LASSO and Ridge models on a test data set.

Mean Squared Error	
LASSO	Ridge
1.608	7.96

3.2.2 Big Signal and Big Noise

Model specifications:

- $\beta = (\underbrace{1, ..., 1}_{700}, \underbrace{0, ..., 0}_{800})^T$
- $p = 1500 > n = 500$
- $X_i \overset{\text{iid}}{\sim} N(0, I)$

For this section we assume that covariates are not correlated, same as before, and there is a big signal, i.e out of 1500 restrictions 700 are true non-zero parameters. Generating the above model and running a regression analysis on the data using LASSO and Ridge we get the results shown in 3.5.

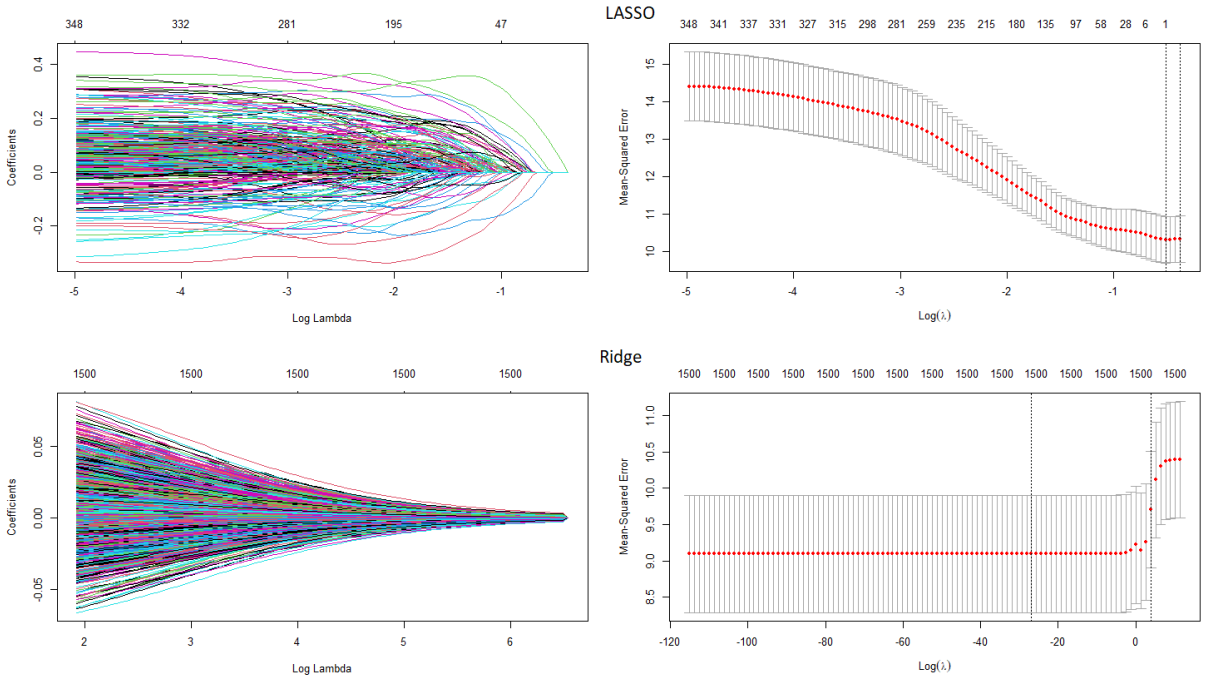

Figure 3.5: Coefficient values against differing log lambda(left side), MSE against varying $\log(\lambda)$(right side)

Through cross-validation we select a λ, which helps us minimise MSE. Comparing the MSE at the chosen λ values for different methods, we see that Ridge performs better than LASSO in the case of Big Signal, which is evident from the table below.

Mean Squared Error	
LASSO	Ridge
9.285	8.623

3.2.3 Correlated Explanatory Variables

Model specifications:

- $\beta = (10, 10, 5, 5, \underbrace{1, ..., 1}_{10}, \underbrace{0, ..., 0}_{36})^T$
- $p = 50$
- $n = 100$
- Correlated Variables:$Cov(X)_{ij} = (0.7)^{|i-j|}$

In this simulation we assume that covariates are correlated, as opposed to the previous cases. Generating the above model and running a regression analysis on the data using LASSO and Ridge we get the results shown in 3.6.

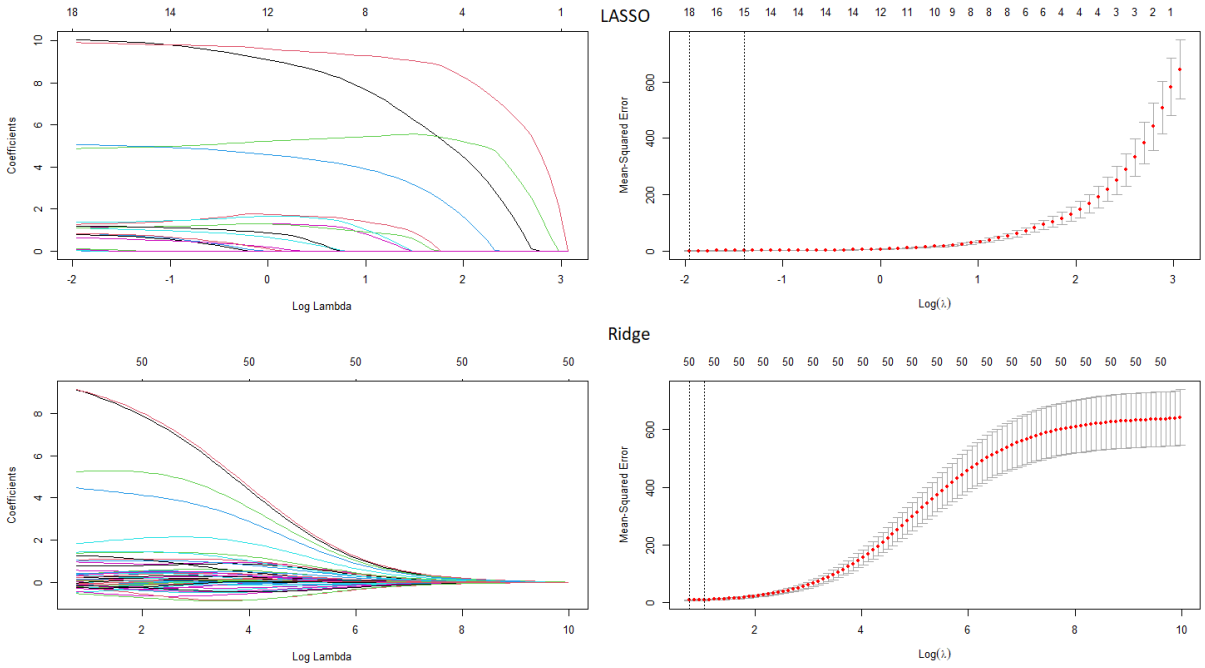

Figure 3.6: Coefficient values against differing log lambda(left side), MSE against varying $\log(\lambda)$(right side)

We see from the above graphs that LASSO does the best job in variable selection considering the number of true non-zero parameters and the data generating setup, while Ridge fails to shrink true zero-valued parameters to zero. Again we turn to MSE calculated on test data for measuring the accuracy of both selected models. As expected, LASSO has a smaller value than Ridge.

Mean Squared Error	
LASSO	Ridge
1.895	7.251

3.3 Distribution of Weights in Elastic Net

3.3.1 Small Signal and a Lot of Noise

We saw that when we have Small Signal data set LASSO performs better. In order to see the weight distribution between LASSO and Ridge in Elastic Net we plot the following graph.

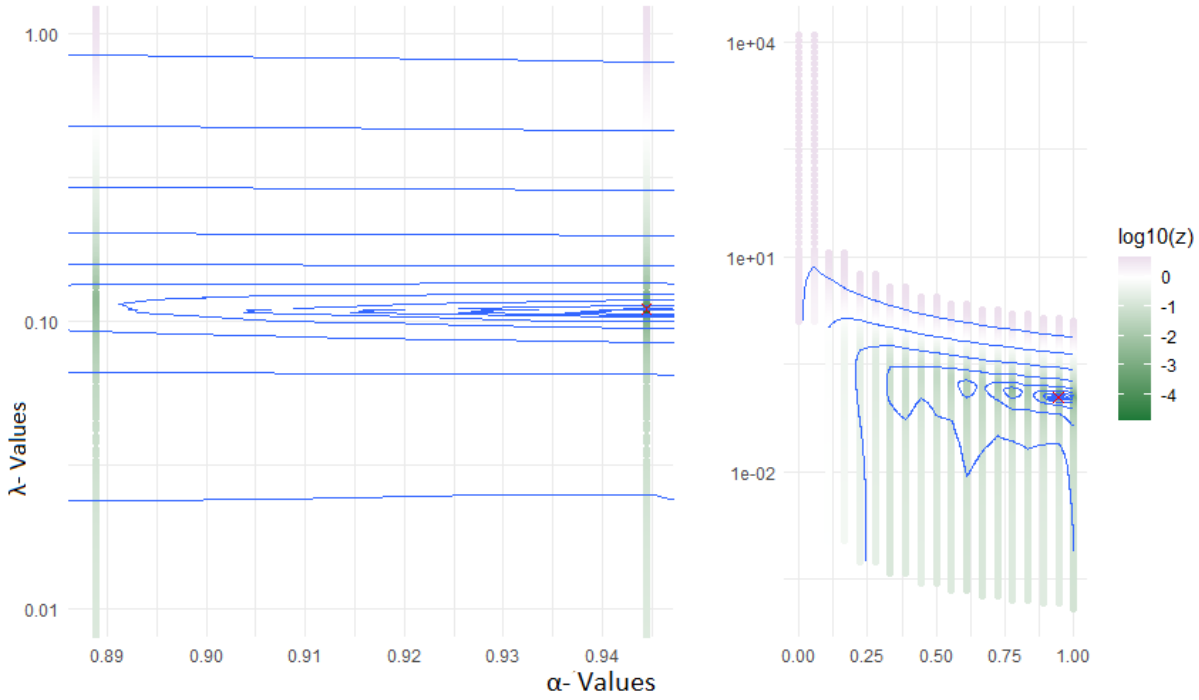

Figure 3.7: Contour Plot of Elastic Net with Small Signal

The weight α is shown on the x-axis and penalty term λ on the y-axis. The higher the α, the more weight elastic net puts on LASSO and vice versus for Ridge. The red dot in the graph indicates the combination of α and λ, which result in the minimum MSE. We see from the graph that more weight is allocated to LASSO, which is because in this particular case LASSO performs better in comparison to Ridge.

3.3.2 Big Signal and Lot of Noise

In the case of big signal we see that Ridge performs better. To illustrate the weight distribution between LASSO and Ridge in the estimated Elastic Net we plot the following graph.

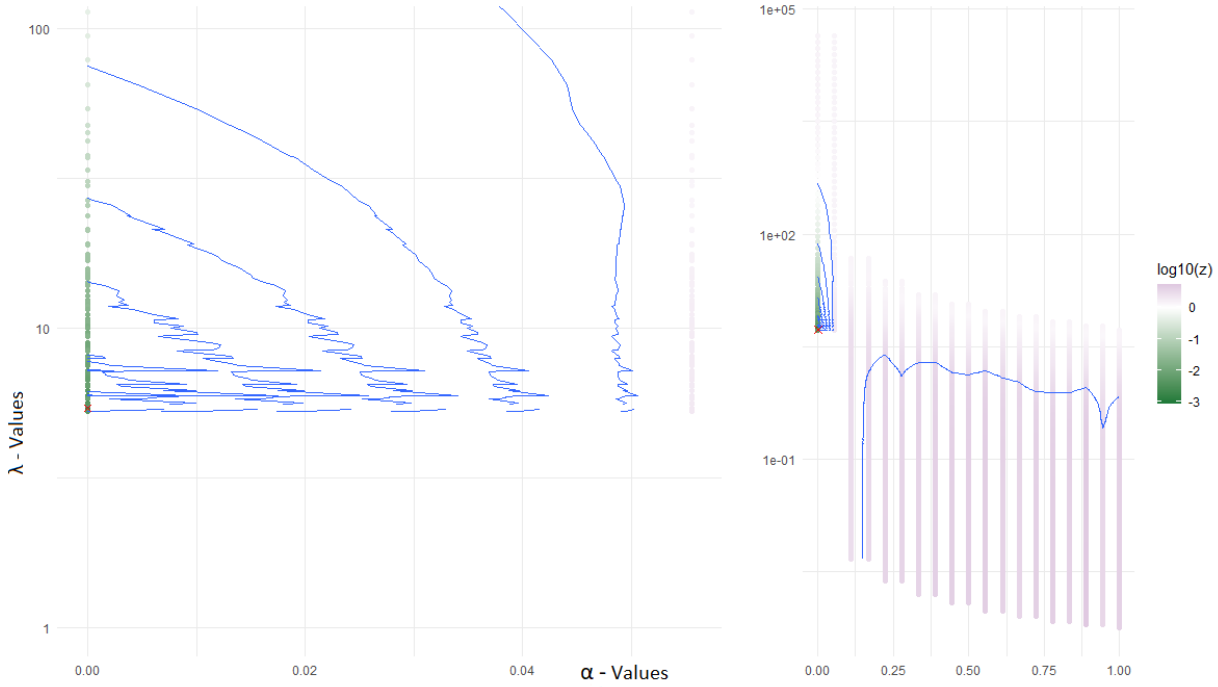

Figure 3.8: Contour Plot of Elastic Net with Big Signal

At the selected α and λ MSE is minimised, and from the graph we see that in this case almost all the weight is allocated to Ridge. This coincides with the fact that in the previous section Ridge performed better than LASSO with Big Signal data set.

3.3.3 Correlated Explanatory Variables

In the previous section we saw that LASSO performs better when explanatory variables are correlated. In order to see the weight distribution between LASSO and Ridge in Elastic Net we plot the following graph, which depicts the selection process for and α aimed at achieving the minimum MSE.

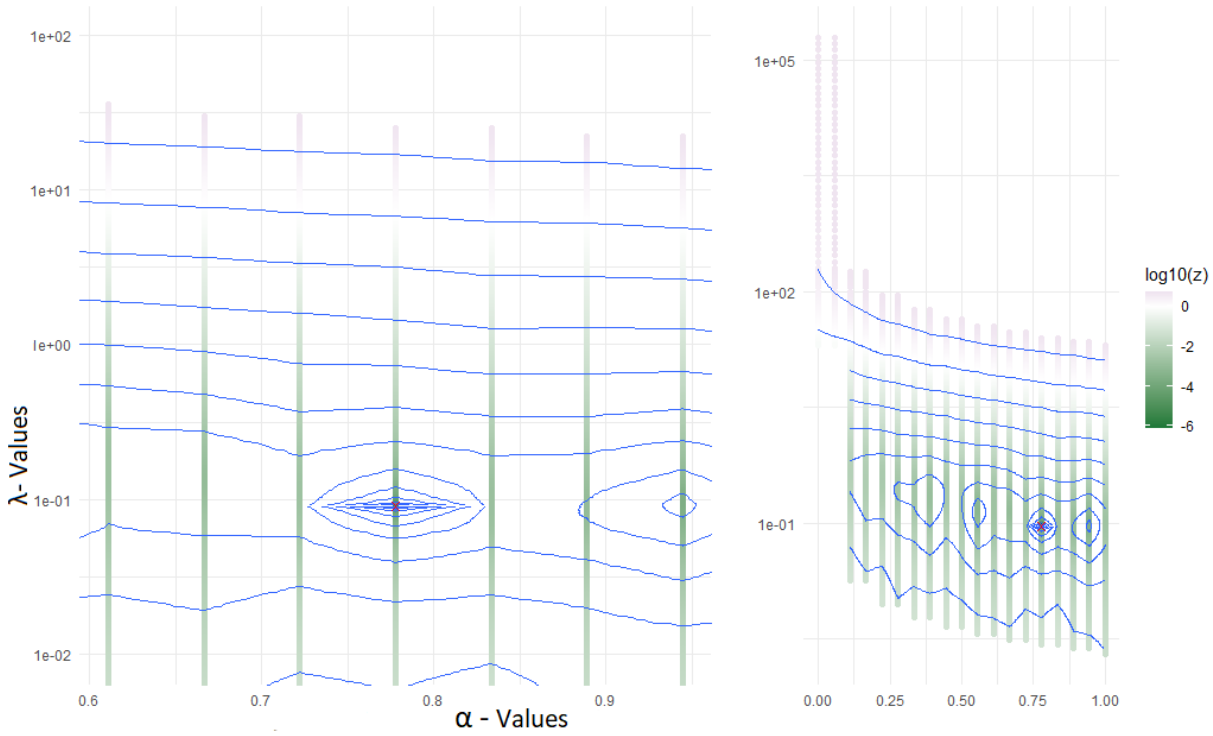

Figure 3.9: Contour Plot of Elastic Net with Correlated Explanatory Variables

Here more weight (α) is allotted to LASSO, since in this particular case it performs better in comparison to Ridge.

3.4 Overall Comparison of LASSO, Ridge and Elastic Net in Varying Signals

Data generating process:

- $\beta = (\underbrace{1, ..., 1}_{k}, \underbrace{0, ..., 0}_{p-k})^T$, k varies from 0 to p
- $p = 1500 > n = 500$
- $X_i \overset{\text{iid}}{\sim} N(0, I)$
- Scaling dependent variable by $\frac{3}{\sqrt{p}}$

Here we try to compare the performance of LASSO, Ridge and Elastic Net using their Mean Squared Error (MSE). From the previous section we see that Elastic Net provides more weight to the best performing method, so we illustrate the weight distribution process for varying signal in the following graph.

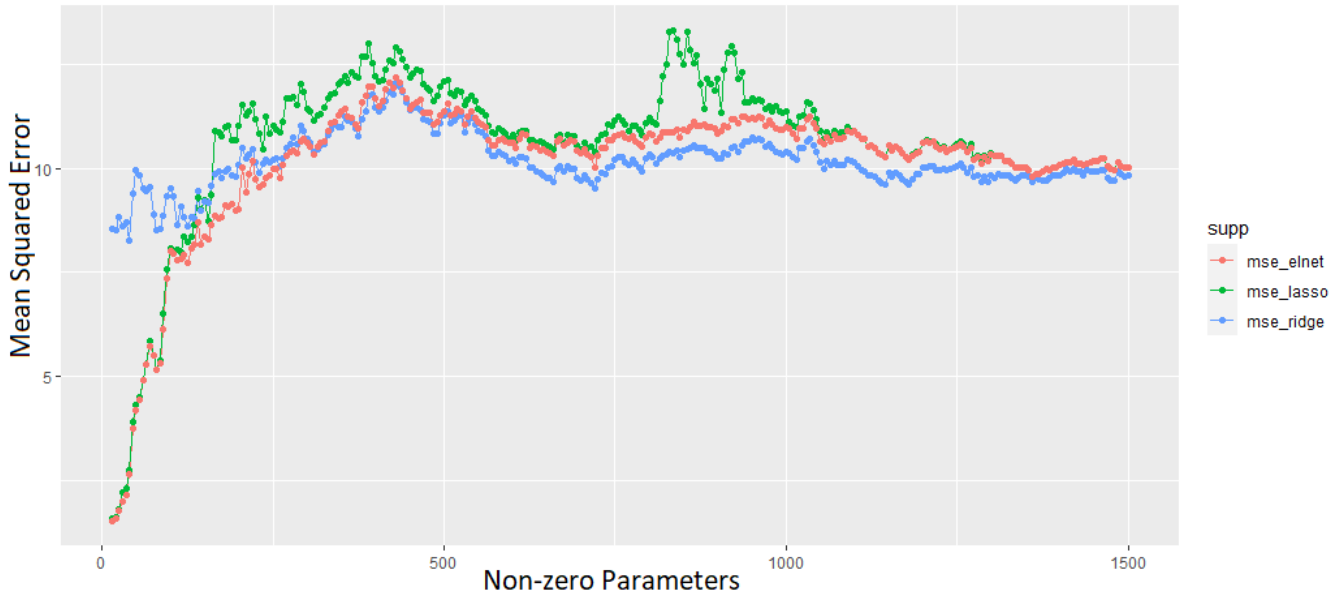

Figure 3.10: MSE of Elastic Net, LASSO and Ridge against Varying Signal

In the figure above we see that when the number of true non-zero parameters is low, LASSO tends to be more accurate than Ridge. However, since Ridge performs better and better as the signal increases, around 100 real non-zero parameters the trend breaks and from then onwards Ridge is more accurate. Elastic net on average tends to have an MSE in between that of LASSO and Ridge, but also a lower MSE than both methods in some cases. The pattern where Elastic Net allocates more weight to the best performing method breaks at very high signals (e.g. 1000 true non-zero parameters) in this particular case.

Chapter 4

Conclusion & Discussion

In the theory section we observed that in the case of high-dimensional setting, when OLS estimators have low bias, but a high variance, regularization is used to reduce variance at the cost of introducing some bias in order to improve the overall prediction accuracy of the model.

We see that in different scenarios (depending on the characteristics of data) different methods, i.e. LASSO or Ridge, are performing better. We can conclude that neither one is overall better. When we have highly-correlated variables, the Ridge regression shrinks those coefficients towards one another, while LASSO generally picks one over the other. In this case the accuracy prediction depends on other characteristics of the data. In general, LASSO tends to do better if there is a small number of significant parameters and all others are close to zero. While using LASSO, it is important to know how LASSO selects if there are differential variances or some of the covariates are correlated with one another.

Ridge tends to work well if there are many non-zero parameters of smaller values.

Elastic net is a compromise between the two that attempts to shrink and do a sparse selection simultaneously. It is a weighted average of both the penalties of LASSO and ridge regression, where we assigned weights using cross-validation. If we have no information about the data, then we can run the elastic net and after choosing the optimal α value, the best-performing method between LASSO and Ridge shall be used.

References

[1] A. Belloni, V. Chernozhukov, and C. Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28:29–50, 2014.

[2] P. Buhlmann and van de Geer. *Statistics for high-dimensional data- Methods, theory and applications*. Springer Series in Statistics Springer, Heidelberg, 2011.

[3] N. Gauraha. Introduction to the lasso. *Reson*, 23:439–464, 2018.

[4] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, 2009.

[5] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[6] A. Hoerl and R. Kennard. *Ridge regression: biased estimation for nonorthogonal problems.* Technometrics, 2000.

[7] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, page 267–88, 1996.

[8] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67:301–320, 2005.

Appendix

Some Other Important Results

Data generating process:

- $\beta = (\underbrace{1, ..., 1}_{k}, \underbrace{0, ..., 0}_{p-k})^T$, k varies from 0 to p
- $p = 1500 > n = 500$
- $X_i \overset{\text{iid}}{\sim} N(0, I)$

In the figure 1, we did not try to control the variance of dependent variable which in turn resulted in high Mean Squared Errors.

In the figure below we see that when the number of true non-zero parameters is low, LASSO tends to be more accurate than Ridge. However, since Ridge performs better and better as the signal increases, around 100 real non-zero parameters the trend breaks and from then onwards Ridge is more accurate. Elastic net on average tends to have an MSE in between that of LASSO and Ridge, but also a lower MSE than both methods in some cases.

Figure 1: MSE of Elastic Net, LASSO and Ridge against Varying Signals

In the Figure 2, we try to reduce the variance by scaling the dependent variable by $\frac{5}{\sqrt{p}}$, which in turn resulted in lower MSE.

We see that LASSO performs better with lower signal. However, as the signal increases, Ridge performs better and better. Considering that DGP behind the data, we know that the true coefficient values are close to zero, which enables Ridge to outperform LASSO quicker in comparison to the previous case.

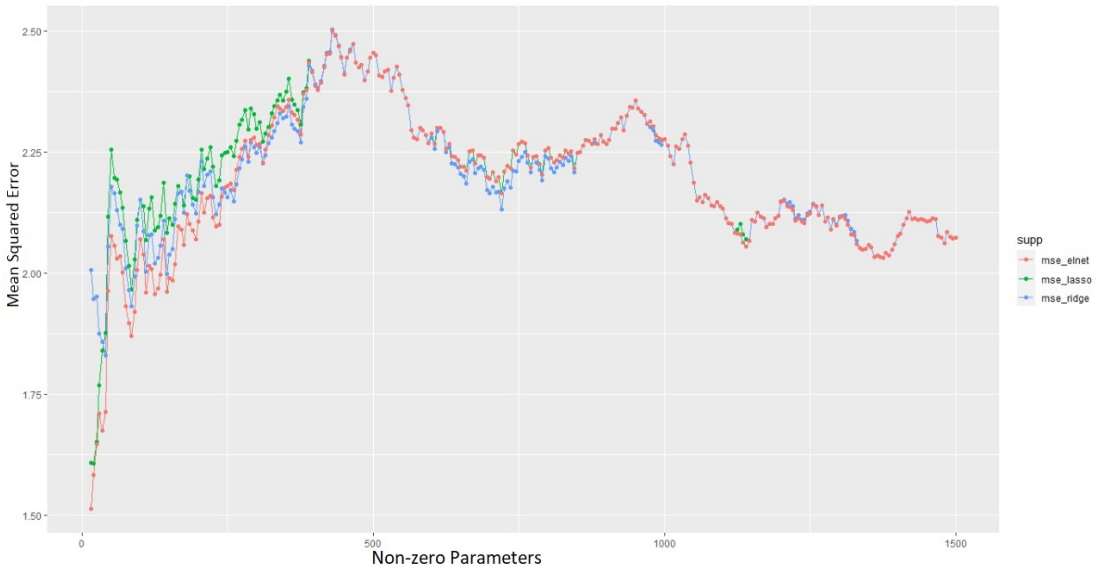

Figure 2: MSE of Elastic Net, LASSO and Ridge against Varying Signals