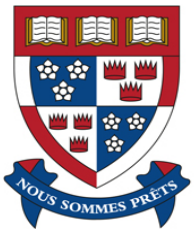

Simon Fraser University
Faculty of Statistics & Actuarial Science
Burnaby, British Columbia

The LASSO

Presented By:

Zubia Mansoor

Contents

1. Motivation

The principle of Occam's Razor states that among many possible explanations for a phenomenon, the simplest is best. Extending this to the regression framework, the goal is to look for the most parsimonious model that best fits the data. There are a number of lens from which you can view this problem; I present the basics of regression shrinkage and selection through the LASSO.

In this section, I briefly review ordinary least squares (OLS), subset selection and ridge regression and highlight certain limitations. Following this discussion, I then introduce lasso in Section 2 as a method of estimation in linear models that overcomes these shortcomings. Further into the section, I introduce the model for the lasso and elaborate on estimation and inference procedures, finishing with a comparison with the standard techniques of subset selection and ridge regression. The lasso idea can be extended to several statistical settings: here I restrict them to elastic net, generalized regression and tree-based methods in Section 3. Finally, some conclusions are provided in Section 4 to summarize the report.

1.1 Ordinary Least Squares

Let us first take a step back and begin with the usual regression approach. As we are familiar, the OLS estimates for $\beta_j's$ are obtained by minimizing the residual sum of squared errors. In other words, we minimize the sum of the squares of the differences between the observed response variable Y, and the fitted values $\hat{Y}$.

Consider a linear model: $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. The OLS coefficients $\hat{\beta}_{OLS}$ are estimated by minimizing the following objective function:

$$S = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 = ||y - X\beta||^2 \tag{1.1}$$

The solution which minimizes the above objective function (1.1) is given by:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y \tag{1.2}$$

Now, the OLS theory has been around for quite some time and has proven to be very useful. However, in certain cases, we are not satisfied with the least square estimates. For instance, suppose we have a model with a large number of predictors. We might have only a few predictors contributing to the effect on the response, but that knowledge is unknown to us. What this translates to is that our coefficient matrix may be sparse, meaning many $\beta_j's$ are exactly 0. Clearly, OLS is not quite helpful in such situations. Hence, there is scope for improvement upon the OLS estimates. Moreover, the criteria for evaluating the quality of a model will differ according to the circumstances. Typically, the following two aspects are important.

Prediction Accuracy: We find that least squares estimates often have low bias but high variance. A way of improving the prediction accuracy is by *regularisation* (shrinking some coefficients) or *sparsity* (setting some coefficients to zero). Since OLS estimates have low bias, shrinking or setting coefficients to 0 would introduce bias in our model. In essence, we increase the bias in our model to reduce the variance of the predicted values, with the hope to achieve better accuracy.

Interpretation: It is hard to distinguish the individual predictor effects if we have a large number of them in the model. What we would rather do is determine a smaller subset of a few meaningful predictors that exhibit the strongest effects. This will allow us to interpret our results better.

Up to this point, we have outlined the limitations of the OLS estimator and hence we can acknowledge the need to come up with methods that improve upon these estimates. In statistical literature, penalization has been proposed to improve upon OLS. There is a whole host of such techniques and we will focus on the standard methods of *subset selection* for dealing with *variable selection* and *ridge regression* to perform *regularisation*.

1.2 Subset Selection

Suppose we have a model with a large number of predictor variables. A classical method in statistics for selecting parameters in a linear model is subset selection [8]. It evaluates a subset of features as a group and determines its suitability to be included in the final model. Hence, it performs variable selection as we retain a subset of the variables and discard the rest from the model. Thereafter, least squares regression is performed to get the estimated coefficients of these retained variables. Thus, it creates *sparsity* by completely removing some predictors from the model but it still lacks in certain aspects. We elaborate on these points below.

More Interpretability: It creates more interpretable models by retaining a subset of the predictors from the original model, so eventually we end up with fewer predictors in the model.

Low Prediction Accuracy: Being a discrete process, subset selection exhibits high variability - small changes in the data can lead to different models being produced each time and this can, in turn, lead to a decrease in overall prediction accuracy.

The key takeaway is that although subset selection produces a sparse model, it is an extremely variable method. In this context, we introduce ridge regression as a more stable method of estimation. It is a shrinkage method that is more continuous, and does not suffer as much from high variability.

1.3 Ridge Regression

Ridge regression [8] shrinks the regression coefficients by imposing a budget or upper bound on their size. In essence, it is minimizing a constrained residual sum of squares given by:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq t \tag{1.3}$$

where, $t \geq 0$ is a complexity parameter that controls the amount of penalty imposed on the model parameters.

As we increase the penalty on the coefficients, more and more of them start shrinking towards 0. Thus, ridge regression *regularises* the coefficient estimates towards 0. However, ridge regression cannot produce a parsimonious model, for it always includes all the predictors in the model. These aspects of ridge regression are summarised as follows:

More Stability: As it is a continuous process that shrinks coefficients towards 0, it exhibits greater stability compared to subset selection. Hence, it is reasonable to say that it is negligibly affected by small changes in the data.

Less Interpretability: Since it does not actually set any coefficients exactly to zero, we retain the same number of predictors as in our original model. So, it does not necessarily provide us with easily interpretable models.

As we end this section, one can imagine that we would be interested in a more sophisticated method that is better equipped to handle high-dimensional data as well as one that overcomes the shortcomings of the aforementioned techniques. To that extent, we motivate the use of the *lasso* that does both *continuous shrinkage* and *variable selection* simultaneously. The takeaway thus far is that we would like a sparse estimator that first detects relevant variables and then estimates their coefficients, all while maintaining a high prediction accuracy. While this will be formalized later, we should keep in mind that variable selection coupled with regularisation is what is at the core of this motivation.

2. The LASSO

2.1 Definition

Robert Tibshirani [9] proposed a new regression analysis technique, called the *lasso*, for 'least absolute shrinkage and selection operator'. The lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Due to the form of this constraint, lasso is able to shrink some coefficients towards 0 and set others exactly to 0. Thus, it performs both *variable selection* and *regularization* in order to enhance the prediction accuracy and interpretability of the produced model. In this way, it retains the favourable properties of both subset selection and ridge regression. Originally formulated in geophysics literature in 1986, it was later popularized in 1996 by Tibshirani, who introduced the term and provided better understanding into its performance.

We now introduce the model for the lasso that covers the basics of the model structure, estimation and inference procedures and ends on a note of comparison with the methods of subset selection and ridge regression.

2.2 Model

2.2.1 Model Assumptions

We begin with the usual regression framework. 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. We assume that either the observations are independent or $y_i's$ are conditionally independent given the $x_{ij}'s$. Futhermore, we assume $x_{ij}'s$ are standardized so that $\sum_{i=1}^{N} x_{ij}/N = 0$ *and* $\sum_{i=1}^{N} x_{ij}^2/N = 1$. OLS estimates are equivariant under the scaling of the inputs (multiplying X by a constant c leads to scaled least squares coefficient estimates by a factor of $1/c$). However, the LASSO solutions are not scale equivariant and therefore we need to standardize the predictors or bring them to the same scale before solving the objective function for the lasso. This prevents penalizing some coefficients more than the others and precludes the influence of large, unscaled values of β. Note that the design matrix need not be of full column rank. In other words,

the lasso is applicable even if the predictors are correlated. The lasso solution is unique when $rank(X) = p$, because the criterion is strictly convex. However, it is not strictly convex when $rank(X) < p$, and so there can be multiple minimizers of the lasso criterion. When the number of variables exceeds the number of observations ($p > n$), we must have $rank(X) < p$.

2.2.2 Model Forms

The lasso is a shrinkage method like ridge regression, with a subtle yet key distinction in terms of the penalty imposed. Here, we will outline the two forms in which the lasso problem can be framed.

Constrained Form:

The lasso estimate is defined by:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t \tag{2.1}$$

The lasso minimizes the residual sum of squares as in OLS regression given by (1.1), with an additional penalty term on the sum of the absolute values of the model parameters. The above objective function is also quite similar to the ridge regression problem given by (1.3): the L_2 constraint in ridge is replaced by the L_1 constraint in lasso. It is due to the L_1 constraint that lasso sets some coefficients exactly to 0 and is able to perform variable selection.

Computing the lasso solution to (2.1) is a quadratic programming problem with linear inequality constraints. The L_1 constraint gives a solution which is non-linear in the y_i, so there is no closed form expression as in ridge regression. However, there exists some efficient and stable algorithms for computing the entire lasso path as t is varied, with the same computational cost as for ridge. We will talk about one such algorithm *least angle regression* in Subsection (2.3.2)

The *tuning parameter* $t \geq 0$ controls the amount of shrinkage that is applied to the estimates. To get a better understanding of this, suppose we have $\hat{\beta}_j^0$ as the full least squares estimates and let $t_0 = \sum_j |\hat{\beta}_j^0|$. Then for values of $t < t_0$, lasso solutions will start shrinking the coeffcients

towards 0, and some will be exactly set to 0. For example, if $t = t_0/2$, then the absolute sum becomes halved and it translates to setting half the predictors in our model to exactly 0 [9]. There are several methods to estimate t through cross-validation, generalized cross-validation and an analytical unbiased estimate of risk that have been described by Tibshirani [9].

Notice that the intercept β_0 is not included in the penalty term because that would then make the method depend on the origin chosen for Y; that is, adding a constant c to each of the targets y_i would not imply a shift of the predictions by the same amount c. In order to avoid that dependence, we re-parametrize the constant β_0 by centering the predictors: each x_{ij} gets replaced with $x_{ij} - \bar{x}_i$. The intercept β_0 is then estimated by $\bar{y}$ for all t. If we further center the responses, we can fit the model without an intercept. Henceforth, we assume that this centering has been done, so that the input matrix X has p (rather than $p+1$) columns.

Penalized Form:

Equivalently, we can frame the lasso problem given by (2.1) in the *Lagrangian form:*

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.2}$$

The *tuning parameter* $\lambda \geq 0$ controls the amount of penalty applied to the estimates. It is basically the amount of shrinkage, where the coefficient values are shrunk towards a central point which is 0 here. This results in simpler and sparser models which are often easier to interpret compared to high-dimensional models with a large number of parameters. As the value of λ varies, it has a corresponding effect on the coefficients which is presented below:

- $\lambda = 0 \rightarrow$ no parameters are eliminated. The estimate is same as the OLS estimates, 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.

Here we observe a trade-off between bias and variance. As λ increases, bias increases but as λ decreases, variance increases. For instance, a low value of λ would lead to more manageable

number of model predictors and lower bias, but at the expense of higher variance. Also note that λ has a one-to-one correspondence with the upper bound t in (2.1). As $t \to \infty$, the problem is equivalent to performing an OLS regression and the corresponding value of λ becomes 0. Analogously, as t becomes 0, all coefficients get set to 0 and $\lambda \to \infty$.

2.3 Model Estimation

2.3.1 Orthonormal Design Case

We shall first consider the case of an orthonormal design matrix where the predictors are independent. This will help us to gain insights about the nature of the shrinkage carried out by lasso. Let X be the $n \times p$ design matrix with ij^{th} entry x_{ij}. In the orthonormal case, $X^T X = I$, where I is the identity matrix.

The solutions to equation (2.1) are shown by Tibshirani in [9] :

$$\hat{\beta}_j = sign(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \lambda)_+ \tag{2.3}$$

where, $\hat{\beta}_j^{0'}s$ are the OLS estimates; λ is a constant chosen by the corresponding technique; sign denotes the sign of its argument (± 1) and x_+ denotes the "positive part" of x, i.e., $max\{0, x\}$.

To get a comparative understanding, we shall also look at the solutions of subset selection and ridge regression in the case of an orthonormal input matrix X. All the three procedures have explicit solutions here. Each method applies a transformation to the OLS estimates $\hat{\beta}_j^0$, as shown in Table 2.1. Below the table, Figure 2.1 shows the geometric form of these estimators.

In the orthonormal design case, best subset selection of size M selects the M largest coefficients in absolute value and sets the rest to 0. For some choice of λ, this is equivalent to setting $\hat{\beta}_j = \hat{\beta}_j^0$ if $|\hat{\beta}_j^0| > \lambda$ and to 0, otherwise. In other words, it discards all variables with coefficients smaller than the M^{th} largest. Ridge regression scales the coefficient by a constant factor, which is why we observe the ridge solutions to be a little tilted away from the 45° line in Figure 2.1. LASSO translates by a constant factor, truncating at 0. So, we get zero-valued coefficients if $|\hat{\beta}_j^0| < \lambda$. For all other values, we get scaled coefficients maintaining the same signs as OLS.

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j^0 \cdot I(\lvert\hat{\beta}_j^0\rvert \geq \lvert\hat{\beta}_{(M)}^0\rvert)$
Ridge	$\hat{\beta}_j^0/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j^0)(\lvert\hat{\beta}_j^0\rvert - \lambda)_+$

Table 2.1: Estimators of β_j in case of orthonormal columns of X.

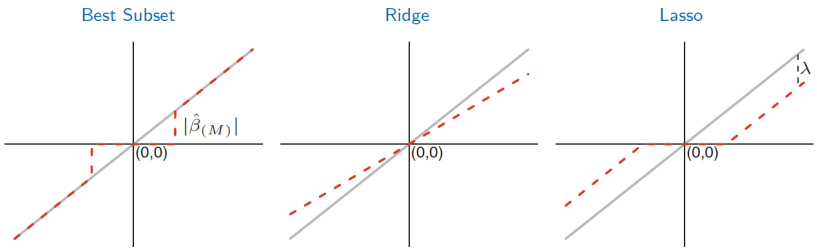

Figure 2.1: This image shows the form of the solutions for subset selection, ridge regression and lasso in case of orthonormal columns of X denoted by the broken red lines. The 45° line in gray represents the unrestricted estimates, that is, the OLS estimates.

2.3.2 General case

In the non-orthogonal case, there is no closed form solution for the lasso, but there are efficient numerical algorithms available to get the entire path of solutions. Now, we talk about one such algorithm *least angle regression*, a variant of which yields all lasso solutions.

Least Angle Regression: Algorithm for the lasso path

A simple modification of the least angle regression (LARS) [1] provides an extremely efficient algorithm for computing the entire lasso path. It continues to provide lasso coefficients along the path, and the final solution highlights the fact that a lasso fit can have no more than $N-1$ (mean centered) variables with non-zero coefficients as is evident by Algorithm 1.

In order to demonstrate the LARS algorithm, we shall display a real-world example for computing the lasso path. Suppose we are interested in the level of prostate-specific antigen (PSA), elevated in men who have prostate cancer. We have the response variable defined as $y_i = \log(PSA)$ and the predictors are x_{ij}, measurements on a man and his prostate. We have data on 97 men with prostate cancer and 8 clinical predictors ($n = 97, p = 8$). We might want to

derive a linear model using only a few of the 8 predictors to predict the level of PSA? Below, we implement lasso for variable slection using LARS algorithm. We observe that the predictors *lcavol, lweight and svi* are retained in the final model as is evident in Figure 2.2 [8].

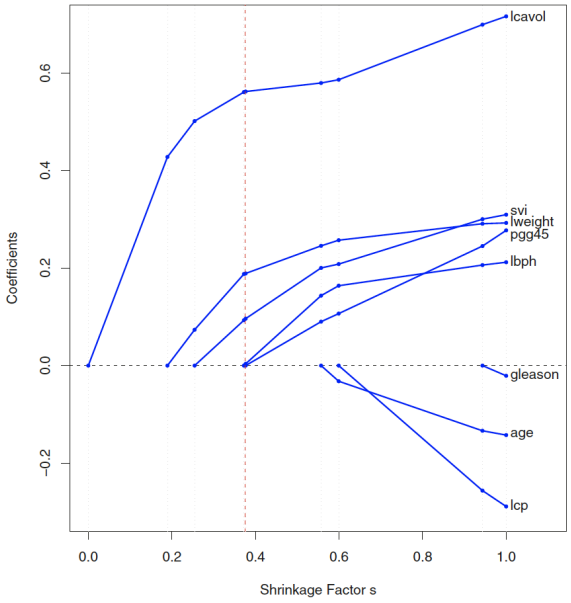

Figure 2.2: This image shows the profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. The profiles are piece-wise linear, and so are computed only at the points displayed.

Algorithm 1 *Least Angle Regression*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}, \beta_1, \beta_2, \ldots, \beta_p = 0$..
2. Find the predictor x_j most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $< x_j, r >$, until some other competitor x_k has as much correlation with the current residual as does x_j.
4. Move β_j and β_k in the direction of the joint least squares coefficient of the current residual on (x_j, x_k), until some other competitor x_l has as much correlation with the current residual.
5. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.
6. Continue in this way until all p predictors have been entered. After $min(N-1, p)$ steps, we arrive at the full least-squares solution.

Geometric Interpretation:

So far, we have discussed that the lasso sets coefficients exactly to zero, while ridge regression does not. We also briefly touched upon the fact that this was attributable to the shape of the constraint boundaries in the two cases. Both the procedures may seem superficially similar as they minimize the same objective function but the key difference lies in the constraint each method employs and we shall examine this graphically in this section.

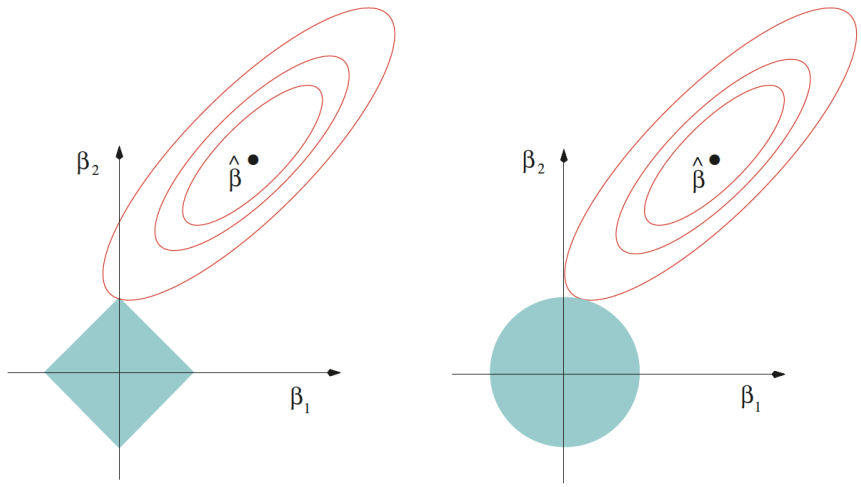

Figure 2.3: This image shows the estimation of coefficients in a two-predictor case for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions and the red ellipses are the contours of the least squares error function.

In Figure 2.3, we observe the constraint region for the lasso and ridge regression for a two-predictor case. In case of lasso, it is the diamond $|\beta_1| + |\beta_2| \leq t$ so that it has corners lying on the axes. Whereas for ridge, the constraint region is the disk $\beta_1^2 + \beta_2^2 \leq t$, which is rotationally invariant and, therefore, has no corners. The residual sum of squares has the geometric form of elliptical contours centered at the least square estimates of the full model. For both the methods, the solution corresponds to the point where the contours *first* meet the constraint region. Recall that the goal is to minimize the residual sum of squares, so we look for its minimum value (corresponding to the innermost level of the contours) that satisfies the constraints.

For lasso, the solution corresponds to the first place where the contours touch the diamond. Thus, if this solution occurs at a corner, we get zero-valued coefficients. Since the ridge constraint

has no corners for the contours to hit, it is not actually possible to set the coefficients to 0. In higher dimensions ($p > 2$), the diamond becomes a rhomboid with several corners to hit. This increases the chance for more and more parameters to be set to 0. As for ridge, the circle transforms into a hypersphere and the objective function can hit the constraint region in more arbitrary places on the sphere, so we do not end up with zero-valued coefficients.

The key takeaway here is that a convex region is more likely to encounter a corner if it lies tangent to a boundary, which then converts some components of β to being identically zero. Contrarily for a hypersphere, the convex region is as likely to contact a point on the boundary at which some β components are zero as the ones for which none of the $\beta's$ are; this is because they are indistinguishable from each other.

Let us consider the L_q norm defined by: $||\beta||_q = (\sum_{i=1}^{n} |\beta_i|^q)^{1/q}$. It is important to note that $q = 1$ is the smallest value of q giving a convex region, which is convenient for optimization. Convex problems are preferred because they are easier to solve; one reason being that any local optimum is also a global optimum. For $q > 1$, $|\beta_j|^q$ is differentiable at 0, and hence does not share the ability of lasso ($q = 1$) for setting coefficients exactly to 0. Therefore, the L_1 norm possesses the desirable properties of convexity and sparsity.

2.4 Inference

It has been shown that the lasso enforces sparsity and selects influential predictors. However, there are still major gaps in making inferences for the lasso model. This is due to the fact that the usual constructs like p-values and confidence intervals do not exist for lasso estimates. As a consequence, these models will always identify a set of the most important predictors even if none or only some of them are significant. So, classical tools cannot be used post-selection because they do not yield valid inferences.

Lockhart [7] developed a method that uses a covariance test statistic approach to compute p-values for parameters obtained from a lasso regression. Meinshausen [6] presents another approach where the data is split into two groups. The lasso is applied to one group, after which

the variables picked by the lasso are used as predictors to obtain p-values from an ordinary least squares regression on the second group. Moreover, there also exists some methods for post-selection inference using the lasso [2]. This method uses the polyhedral lemma to provide an optimal solution for the lasso based on a fixed λ. Using this, they construct selection-adjusted intervals for the parameters of the fitted model.

2.5 Comparison with Subset Selection and Ridge Regression

So far, we have covered the fundamentals of lasso and at every step of the way tried to highlight how it differs from subset selection and ridge regression. The lasso has been shown to emerge as an improvement upon these techniques but it might be useful to know when each of the procedures are appropriate in practice. We talk about this in a more detailed manner in this section. The term *effects* is used to emphasize the impact of a predictor on the response variable.

Small number of large effects: In this case, subset selection performs the best because it is able to make a clear distinction between the predictors who have a strong impact on the response versus the predictors who do not. As a result, it does not suffer as much from variability as it normally would. Contrarily, the lasso does not perform quite as well and ridge regression exhibits a poor performance too. This is due to the fact that we have a few meaningful predictors and shrinking or deleting any of them would only introduce bias in the model.

Small to moderate number of moderate-sized effects: Lasso would be a strong candidate in such a situation, followed by ridge regression and then subset selection. Lasso does a good job here by shrinking some coefficients and setting redundant variables to 0. As for ridge, it would incorporate all the model predictors including those that do not have a strong impact on the response. Moreover, subset selection would be extremely variable in this case as it might exclude some predictors having an effect on the response.

Large number of small effects: Ridge regression does best by a good margin, followed by the lasso and then subset selection. If there is a large number of predictors affecting the outcome, ridge would a good idea because it would incorporate the effect of each of those predictors

through regularisation. However, lasso would perform variable selection and end up with high bias in case of deletion of important variables. Once again, subset selection would exhibit poor performance due to extreme variability in terms of the model.

At the end of this section, we have covered the basics of the lasso model and briefly talked about estimation and inference procedures. Additionally, we compared its performance with the standard methods under different circumstances. At this point, we can take a step forward and consider some extensions of the lasso. To that extent, I have described three such extensions in the forthcoming section.

3. Extensions

3.1 Elastic Net

We have established the fact the lasso is an improvement upon subset selection and ridge regression in the sense that it performs both regularsiation and variable selection. However, there are some limitations to the lasso which are outlined as follows:

- In high-dimensional data settings ($p > n$), the lasso selects at most n variables before it saturates. Thus, the number of variables selected is limited by the number of observations.
- In case of highly correlated variables, the lasso tends to select one variable from the group and disregards the impact of the other predictors in the group, so it fails to perform grouped selection.

To address the above shortcomings of the lasso, Zou and Hastie [10] proposed the *Elastic Net* which is a more refined method of regularization and variable selection. It employs a penalty function that is a linear combination of the L_1 and L_2 penalties of the lasso and ridge methods. The estimates from the elastic net method are defined by:

$$\hat{\beta}_{EN} = \arg\min_{\beta} \left\{ ||y - X\beta||^2 + \lambda_2 ||\beta||^2 + \lambda_1 ||\beta||_1 \right\} \tag{3.1}$$

Elastic net enjoys the property of sparsity due to the L_1 part of the penalty. On the other hand, the quadratic part of the penalty serves three purposes. First, it elimiates the bound on the number of selected variables. Hence, it is particularly useful when the number of predictors exceed the number of observations by a large margin. Secondly, unlike lasso, it encourages grouping effect so that strongly correlated predictors tend to be in or out of the model together. Finally, it provides a more stable path for the L_1 regularization.

3.2 Generalized Regression Models

Tibshirani [9] briefly talked about the application to generalized regression models. Here, one takes a model indexed by a vector parameter β. The estimation is then carried out by

maximization of a function $l(\beta)$; in most cases it is the log-likelihood function. However, one can alternatively use some other measure of fit. In order to perform the lasso, we maximize $l(\beta)$ under the L_1 constraint such that $\sum_j |\beta_j| \leq t$. This maximization can be possibly executed by a general (non-quadratic) programming procedure. But Tibshirani [9] instead considers different models for which a quadratic approximation to $l(\beta)$ leads to an iteratively reweighted least squares (IRLS) procedure for computation of β. This is equivalent to using the numerical algorithm of Newton-Raphson. Using this approach, we can solve the constrained problem by iterative application of the lasso algorithm, within an IRLS loop. This procedure shows promise in its performance, however convergence is not generally ensured.

3.3 Tree-based methods

LeBlanc and Tibshirani [4] extend the concept of lasso for tree-based methods. The central idea was to incorporate lasso in regression trees to shrink them instead of pruning a large tree using classification and regression tree (CART) algorithm. We follow a similar methodology to implement lasso as described earlier to obtain predictions in this case. Essentially, the sum of the absolute values of the parameters representing the effects of splits are constrained to be less than a constant. The parameters here correspond to the mean contrasts at each node. This method based on the lasso leads to both shrinking at each node and pruning of branches in the tree. In some cases, it gives more accurate predictions than cost-complexity pruning used in the CART approach and also produces more interpretable subtrees.

Ideally, we would like a tree-based model that yields accurate predictions and is simple to interpret. Le Blanc and Tibshirani [4] aim to achieve this by doing the following:

- Specifying feature variables x_{ij} to yield coefficients which can be interpreted easily in a tree-based model.
- Incorporate constraints that lead to pruning of trees.

Traditional algorithms for growing tree-based models include forward and backward model construction steps. Applying the idea of lasso for simplifying trees can be viewed as a smoother type of pruning.

4. Conclusions

In modern data analysis tasks, it has become increasingly important to perform variable selection in order to pick the best-fitting model among a plethora of options. Previously, we only had a few carefully chosen predictors for each observation. In contrast to this, nowadays any variable that might possibly affect the response is included as a predictor. In light of this, *the least absolute shrinkage and selection operator* has proven to be a valuable tool for the purpose of estimating the coefficients and creating more interpretable models, especially in the high-dimensional setting. It has positively affected countless other fields and its versatility is latently evident by the growing literature in this field.

In this short report, I merely presented the fundamentals of the model; more specifically, I aimed to motivate the approach with a need to adapt and innovate over the limitations in the existing methods of variable selection. Stepping back into the ordinary least squares framework naturally gave way to the bedrock of the methodology and substantially explained how the lasso estimator behaves, including its relationship to subset selection and ridge regression. This was followed by an overview of model estimation and inference for the purpose of achieving a general understanding of how lasso operates. This set in motion the tools needed to put this model to use in the way of applications to real-life data sets. There are several generalizations of the lasso that offer insights into its performance and ways of improvement and we briefly talked about elastic net, generalized regression and tree-based methods.

In the ever expanding literature of model selection techniques, the lasso has carved out a fairly prominent niche - particularly for high dimensional sparse data sets. What was presented here is just a snippet of an expansive method for which there are a seemingly endless number of avenues to go down. Whether that be overcoming efficiency problems using BASIL [3] or using a variant of the original version such as grouped lasso [5], the lasso has shown potential and prowess. All of this and more makes the lasso a fascinating and growing field to do future research on!

References

[1] Iain Johnstone Bradley Efron Trevor Hastie and Robert Tibshirani. "Least Angle Regression". In: *The Annals of Statistics, Vol. 32, No. 2, 407–499* (2004).

[2] Yuekai Sun Jason D. Lee Dennis L. Sun and Jonathan E. Taylor. "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics, Vol. 44, No. 3, 907–927* (2016).

[3] Yosuke Tanigawa Matthew Aguirre Robert Tibshirani Manuel A. Rivas Trevor Hastie Junyang Qian Wenfei Du. "A Fast and Flexible Algorithm for Solving the Lasso in Large-scale and Ultrahigh-dimensional Problems". In: *Preprint, bioRχiv, the preprint server for biology* (2019).

[4] Michael LeBlanc and Robert Tibshirani. "Monotone Shrinkage of Trees". In: *Journal of Computational and Graphical Statistics, Vol. 7, No. 4, pp. 417-433* (1998).

[5] Yi Lin Ming Yuan. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society. Series B (statistical Methodology). Wiley. 68 (1): 49–67.* (2006).

[6] Lukas Meier Nicolai Meinshausen and Peter Bühlmann. "P-Values for High-Dimensional Regression". In: *Journal of the American Statistical Association Vol. 104, No. 488, pp. 1671-1681* (2009).

[7] Ryan Tibshirani Robert Tibshirani Richard Lockhart Jonathan Taylor. "A significance test for the lasso". In: *The Annals of Statistics, Volume 42, Number 2, 413-468* (2014).

[8] R. Tibshirani T. Hastie and J. Friedman. "The Elements of Statistical Learning". In: *Springer Series in Statistics Springer New York Inc., New York, NY, USA* (2001).

[9] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1, pp. 267-288* (1996).

[10] Hui Zou and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 67, No. 2, pp. 301-320* (2005).