STARS

University of Central Florida
STARS

Electronic Theses and Dissertations, 2004-2019

2013

Sparse Ridge Fusion For Linear Regression

Nozad Mahmood
University of Central Florida

Part of the Statistics and Probability Commons

Find similar works at: https://stars.library.ucf.edu/etd

University of Central Florida Libraries http://library.ucf.edu

STARS Citation

University of Central Florida

STARS
Showcase of Text, Archives, Research & Scholarship

SPARSE RIDGE FUSION FOR
LINEAR REGRESSION

by

NOZAD HUSSEIN MAHMOOD
B.S. University of Salahaddin-Hawler, 2003

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Statistics
in the College of Science
at the University of Central Florida
Orlando, Florida

Fall Term
2013

Major Professor: Edgard Maboudou

ABSTRACT

For a linear regression, the traditional technique deals with a case where the number of observations n more than the number of predictor variables p (n>p). In the case n<p, the classical method fails to estimate the coefficients. A solution of this problem in the case of correlated predictors is provided in this thesis. A new regularization and variable selection is proposed under the name of Sparse Ridge Fusion (SRF). In the case of highly correlated predictor , the simulated examples and a real data show that the SRF always outperforms the lasso, elastic net, and the S-Lasso, and the results show that the SRF selects more predictor variables than the sample size n while the maximum selected variables by lasso is n size.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Regression analysis was first developed by Sir Francis Galton during the late 19th century. Galton had observed the relation between heights of parents and offspring, and he noted that the heights of children of both tall and short parents appeared to regress towards the mediocre point (mean of the group) [11]. Regression analysis is one of the most commonly used techniques for analyzing multi factor data. Its wide resumption and utility result from the conceptually logical process of using an equation to express the relationship between a response variable and one or more predictor variables. Because of elegant basic mathematics and statistically advanced theory, the regression analysis is also interesting theoretically [23].

There are different types of regression models, the linear regression and non-linear regression model. In the linear regression model, the parameters are linear because no parameter appears as an exponent or is multiplied or divided by another parameter [24]. The model which includes only one predictor variable is called simple linear regression, and the model with more than one predictor variables is a multi linear regression model. An instance of nonlinear regression is logistic regression, proposed by Berkson in 1944 with the introduction of the $logit$ model [1]. Logistic regression is used for the situations where the response variable is a binary value (0 or 1). This method yields a prediction equation, which is constrained to lie between 0 and 1. Also, some response variables are counts and the two most popular kinds of regression for count variables are Poisson regression and negative binomial regression. Each fits a log-linear model involving both quantitative and categorical predictors. Nonparametric regression

analysis is another kind of regression models that is a regression model without an assumption of linearity. It requires larger sample sizes than regression based on parametric models because the data must supply the model structure as well as the model estimates [9]. Two nonparametric regressions are Kernel regression and Local polynomial regression. The goal of the kernel regression is to obtain efficient predictive method. Local polynomial regression uses weighted least squares regression to generate estimated of a mean function at each point of interest. When the degree of the polynomial is zero, it becomes the kernel regression [12]. The most important challenge in the nonparametric regression fitting is selecting a suitable bandwidth (smoothing parameters). The model should find a balance between the variance and bias in order to get a good fit and leads to the minimization of a mean squared error criterion [21].

The work in this thesis is focused on linear regression. Classical linear regression deals with the case where the number of variables is less than the number of observations (p<n); but when p>n, the classical least squares method cannot be used. Alternative methods based on penalized models are used. Ridge regression, was first published by Hoerl and Robert Kennard in 1970 [16]. The ridge regression penalty ($\|\beta\|_2^2$) shrinks coefficients toward a common value. Lasso regression penalty was introduced by Tabshirani in1996. It is another method to solve the regression problem when p>n. Lasso uses $L_1 - penalty$ ($\|\beta\|_1$) to shrink some coefficients toward zero. If there is a group of highly correlated predictors the lasso picks only one predictor and drops the rest from the model [27]. There are several algorithms to solve the lasso, for instance, the least angle regression (LARS) algorithm, developed by Efron et al. in 2004 [6]. This algorithm is

similar to forward stepwise regression starting from the null model, and picking the predictor variable whose coefficient is most correlated with the residuals at each step [29]. Another algorithm is the coordinate descent, which is shown to be faster than the LARS algorithm [10]. The coordinate descent algorithm was introduced independently by (Friedman at al., 2008) and (Wu and Lange, 2008). It is a path-wise algorithm that can work on very large datasets, and can take advantage of sparsity in the feature set [10, 31]. Another method that performs better than the ridge and lasso is the elastic net, introduced by (Zou and Hastie) in 2005. The elastic net combines both the L_1 and L_2 penalties. It is used for estimation and model reduction [34].The smooth Lasso method was introduced by (Hebiri & Van De Geer) in 2008. It uses the L_1 and fused $\left(\sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2 \right)$ penalties [15]. The local constancy and local linearity are two methods that were suggested by (Hawkins & Maboudou-Tchao) in 2013. The locally constant technique combines the L_2 and $\sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2$ together, and the locally linear combines the L_2 and $\sum_{j=2}^{p-1} (\beta_{j+1} - 2\beta_j + \beta_{j-1})^2$ together [14].

Even though, the previous work gives good results to estimate and select variables, some methods need improvement. For instance, the ridge regression cannot get any interpretable model as it doesn't set any coefficient to zero. The lasso penalty also has some problems. When n>>p, the maximum variable selection of the lasso is n, and with a highly correlated dataset, it only selects one variable and drops the others. Sometimes, the elastic net penalty cannot give good results for the linear and logistic

regression models, and it breaks down on model selection consistency when p>>n [4, 17, 35].

In this thesis, a new penalization method is introduced and called Sparse Ridge Fusion (SRF). The technique of the SRF is mixture of the L_1-lasso penalty ($\sum_{j=1}^{p} |\beta_j|$) and L_2-$\sum_{j=2}^{p-1}(\beta_{j+1} - 2\beta_j + \beta_{j-1})^2$.

The rest of the thesis is organized as follows: In chapter two, the existing methods will be reviewed, chapter three shows the solutions of the S-Lasso by using the coordinate descents algorithm and discusses the SRF, chapter four explains the computations of S-Lasso and SRF, chapter five presents some simulated examples and a dataset, and chapter six shows the conclusion of the thesis.

CHAPTER TOW: LITERATURE REVIEW

In this chapter, the principles and techniques of some previous penalized methods are presented.

2.1. The principle of the linear regression model

The regression model is a statistical relation that gives two main points:

1. The regression function of Y on X represents the relationship of the mean of the probability distribution of Y as a function of X; it captures the notion that Y varies systematically as a function of X

2. The error term represents the deviation of Y from the regression function; there is a probability distribution of Y for each level of X that represents the scatter of points around the main direction [18].

When the regression function is linear, the simple linear regression model is written as following equation.

$$y = X\beta + \varepsilon \qquad (2.1)$$

Where:

$\varepsilon \sim N_n\left(0, \sigma^2\, I\right)$

y: is a vector of observation with length n

X: is a design matrix n x p, $X = (x_1, x_2, \ldots, x_n)^T$

β: are coefficients vector of length p, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$

Assume that the response is centered and the predictors are standardized. Consequently:

$$\sum_{i=1}^{n} y_i = 0 \ , \sum_{i=1}^{n} x_{ij} = 0 \ , \ \sum_{i=1}^{n} x_{ij}^2 = 1 \ , \ \forall j \in \{1,2,\dots,p\} \text{ and } \ \forall i \in \{1,2,\dots,n\}$$

Note that, since the predictors are standardized and the response is centered, no intercept has to be estimated.

The usual estimation procedure for the parameter vector $\hat{\beta}$ as a function of X and y is obtained by minimizing sum of the squared errors with respect to β.

$$\hat{\beta} = \ \text{argmin}_\beta (y - X\beta)^T (y - X\beta) \tag{2.2}$$

It turns out that $\hat{\beta}$ has the form if n>p:

$$\hat{\beta} = \ (X^T X)^{-1} X^T y \tag{2.3}$$

The classical least-squares regression cannot use with the dataset that consists of many more predictor variables with pairwise highly correlated than the observations, p>>n [7]. When p >> n, the matrix $(X^T X)$ is not invertible. On the other hand, when the rows of the design matrix X are highly correlated, the $\hat{\beta}$ coefficients are dependent on different x_i .

2.2. Penalized Least Squares

Regularization process for classical regression models are dependent on penalized least squares:

$$PLS(\lambda, \beta) = (y - X\beta)^T (y - X\beta) + p(\lambda, \beta) \tag{2.4}$$

$p(\lambda, \beta)$ is a penalty term and is based on the tuning parameter λ that controls the shrinkage estimates. Where the tuning parameter $\lambda = 0$, the penalty term will not have any impact on the equation (2.4), and the ordinary least squares solution is obtained. Conversely, the larger the penalty applied, the further the estimates are shrunk towards zero. The estimates of the parameter β are acquired by minimizing the equation (2.4) [8].

$$\hat{\beta} = argmin\,\{PLS(\lambda, \beta)\} \qquad\qquad\qquad (2.5)$$

2.2.1. Ridge Regression

Ridge regression, proposed by (Hoerl and Robert Kennard) in 1970, adds an $L_2 - penalty$ ($\|\beta\|_2^2$) term to residual sum of squares function [16].

In this model, $p(\lambda, \beta) = \lambda\,\|\beta\|_2^2$

$$f(\beta) = \|y - X\beta\|_2^2 + \lambda\,\|\beta\|_2^2 \qquad\qquad\qquad (2.6)$$

The ridge coefficients minimize a penalized residual sum of squares.

$$\hat{\beta}_{Ridge} = argmin\,_\beta\|y - X\beta\|_2^2 + \lambda\,\|\beta\|_2^2 \qquad\qquad (2.7)$$

Where:

$$\|y - X\beta\|_2^2 = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 \qquad i = 1,2,\ldots,n \;\; and \;\; j = 1,2,\ldots,p$$

$$\|\beta\|_2^2 = \sum_{j=1}^{p}\beta_j^2 \qquad\qquad\qquad j = 1,2,\ldots,p$$

λ: is the shrinkage parameter which controls the size of the coefficients and amount of regularization. As λ goes up, the amount of shrinkage goes up.

The ridge regression is a shrinkage method, not model selection because the coefficients are shrunken towards zero, but will never become exactly zero [26, 30]. If predictors are very similar, the ridge regression shrinks coefficients toward a common value and tends to give the predictors all equal coefficients. Though the ridge regression method is computationally simple and standard least square can be used to estimate the coefficients, it still cannot get any interpretable model as all the coefficients are still in the model.

2.2.2. Lasso Regression

Least Absolute Shrinkage and Selection Operator (lasso) was proposed by (Tabshirani 1996). Lasso methods are mostly used in problem with big datasets, such as genomics [10]. The lasso expected many coefficients to be zero and a small subset to be nonzero. [25]

The lasso problem uses the L_1- penalized least squares criterion to obtain a sparse solution to the optimization problem [27].

$$f(\beta) = \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \qquad (2.8)$$

In the equation (2.8) the penalty term is:

$$p(\lambda, \beta) = \lambda \|\beta\|_1$$

Minimizing the equation (2.8) gives:

$$\hat{\beta}_{Lasso} = argmin_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \qquad (2.9)$$

Where:

$$\|y - X\beta\|_2^2 = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 , \quad i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, p$$

$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$: is L_1-norm penalty

The lasso does both ongoing shrinkage and automatic variable selection altogether. The lasso incorporate the beneficial features of backwards-stepwise selection and ridge regression to equip a sparse, comparatively stable model. Backwards-stepwise selection "starts with the full model and sequentially deletes the predictor that has the least impact on fit" (13).

Though the lasso enables a sparse model, it is unstable with high-dimensional data and cannot select more variables than the sample size before it saturates when $p > n$ [3, 20, 22, 28, 32, 33].

In general, the lasso fails in the two following cases. [34]

a)   When p>>n, the lasso cannot select more than n variables before it saturates. This is a restricted feature for a variable selection method.

b)   If there is a group of highly correlated variables, the lasso only selects one variable from the group and ignores the rest. The lassodoes not care which one is selected.

2.2.2.1. Coordinate Descent for lasso problem.

In this section, the coordinate descent algorithm is used to solve the lasso problem. Each coordinate minimization can be done quickly and the related equations can be updated as it cycles through the variables [10].

9

The lasso problem in equation (2.8) is equivalent to:

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \qquad (2.10)$$

The equation (2.10) can be written as below:

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k - x_{ij}\beta_j\right)^2 + \lambda \sum_{k\neq j}|\hat{\beta}_k| + \lambda|\beta_j| \qquad (2.11)$$

The equation (2.11) is equivalent to the following equation:

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(r_i - x_{ij}\beta_j\right)^2 + \lambda \sum_{k\neq j}|\hat{\beta}_k| + \lambda|\beta_j| \qquad (2.12)$$

Where:

r_i: is a partial residual

$$r_i = y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k$$

Then, by taking the derivative of the equation (2.12) with respect to β_j the solutions of β_j as given by:

$$\frac{df(x)}{d\ \beta_j} = \begin{cases} -\sum_{i=1}^{n} r_i x_{ij} + \beta_j \sum_{i=1}^{n} x_{ij}^2 + \lambda & , \qquad \text{if } \beta_j > 0 \\ -\sum_{i=1}^{n} r_i x_{ij} + \beta_j \sum_{i=1}^{n} x_{ij}^2 - \lambda & , \qquad \text{if } \beta_j < 0 \end{cases}$$

Solving $\frac{df(x)}{d\ \beta_j} = 0$ gives the solutions

$$\beta_j = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i x_{ij} - \lambda}{\sum_{i=1}^{n} x_{ij}^2} & , \qquad \text{if } \beta_j > 0 \\ \dfrac{\sum_{i=1}^{n} r_i x_{ij} + \lambda}{\sum_{i=1}^{n} x_{ij}^2} & , \qquad \text{if } \beta_j < 0 \end{cases}$$

If the response variable is centered and the predictors are standardized, then $\sum_{i=1}^{n} x_{ij}^2 = 1$, and the β_j values become

$$\hat{\beta}_j = \begin{cases} \sum_{i=1}^{n} r_i x_{ij} - \lambda & , & if \ \beta_j > 0 \\ \sum_{i=1}^{n} r_i x_{ij} + \lambda & , & if \ \beta_j < 0 \end{cases} \qquad (2.\,13)$$

By adopting the soft-threshold [5] to write the coordinate descent solution of the lasso problem, equation (2.13)

$$\hat{\beta}_j^{lasso} = S(z, \lambda) \qquad (2.\,14)$$

Where S is the soft threshold operator, and is defined as the bellow:

$$\hat{\beta}_j^{lasso} = \begin{cases} z - \lambda & , & if \ z > 0 \ and \ \lambda < |z| \\ z + \lambda & , & if \ z < 0 \ and \ \lambda < |z| \\ 0 & , & if \qquad \lambda \geq |z| \end{cases} \qquad (2.\,15)$$

Where:

$$z = \sum_{i=1}^{n} r_i x_{ij}$$

2.2.3. Elastic Net Penalty

The elastic net was introduced by (Zou and Hastie 2005). Like lasso, it does both automatic estimation and variable selection of the model altogether. "It is like a stretchable fishing net that retains 'all the big fish' "[34]. Furthermore, the elastic net can select variables more than the sample size. The elastic net penalty is a combination of $L_1 \ and \ L_2$ penalties, and is given by:

$$\hat{\beta}_{enet} = argmin_\beta \ \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \qquad (2.\,16)$$

Where:

$\lambda_1 \|\beta\|_1 = \lambda_1 \sum_{j=1}^{p} |\beta_j|$: is the L_1 (Lasso) penalty

$\lambda_2 \|\beta\|_2^2 = \lambda_2 \sum_{j=1}^{p} \beta_j^2$: is the L_2 (Ridge regression) penalty.

The elastic net penalty method depends on choosing two parameters, λ_1 from the L_1 penalty and λ_2 from L_2 penalty.

Assume that $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$, then the equation (2.15) is equivalent to minimize of:

$$\hat{\beta}_{enet} = argmin_\beta \; \|y - X\beta\|_2^2 \; subjet \; to \; (1-\alpha)\|\beta\|_1 + \alpha \, \|\beta\|_2^2 \leq t \; for \; some \; t$$

(2. 17)

Where:

$(1-\alpha)\|\beta\|_1 + \alpha \, \|\beta\|_2^2$: is the elastic net penalty

For $\alpha = 1$ the elastic net penalty becomes the ridge regression, and for $\alpha = 0$ it becomes the lasso penalty.

2.2.4. Smooth Lasso (S-Lasso)

The smooth lasso (S-Lasso) is first suggested in [15]. It is a mixture of L_1-lasso penalty and L_2-fusion penalty and it is given by:

$$\hat{\beta}_{S-Lasso} = argmin\left\{\frac{1}{2} \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=2}^{p}\left(\beta_j - \beta_{j-1}\right)^2\right\}$$

(2. 18)

Where:

$i = 1, 2, \ldots, n \qquad j = 1, 2, \ldots, p$, and

$\lambda_1 \sum_{j=1}^{p}|\beta_j|$: is the L₁-Lasso penalty which is used for sparsisty in the coefficients

$\lambda_2 \sum_{j=2}^{p}\left(\beta_j - \beta_{j-1}\right)^2$: is the L₂ -Fusion penalty which was introduced in [19].

Smooth lasso penalty tries to stop not only the erratic coefficient, but also coefficients that differ basically from their neighbors [14]. By tuning $(\lambda_1, \lambda_2) \geq 0$ in the equation (2.18), the S-Lasso penalty controls the smoothness of the model. In the first

paper of the S-Lasso, The LARS algorithm solution was used to solve the S-Lasso problem. In this work, the coordinate descent algorithm solution is used to solve S-Lasso problem.

2.2.5. Local Constancy and Local Linearity penalties

The local constancy penalty was proposed by (Hawkins & Maboudou-Tchao) in 2013. It combines both $\|\beta\|_2^2$ and $\sum_{j=2}^{p}(\beta_j - \beta_{j-1})^2$ penalties together and add it to the function of residual sum of square as shown in bellow:

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \underbrace{\lambda_1 \sum_{j=1}^{p}\beta_j^2 + \lambda_2 \sum_{j=2}^{p}(\beta_j - \beta_{j-1})^2}_{penalty\ term} \qquad (2.\ 19)$$

Where:

$$j = 1,\dots,p \quad , \quad i = 1,\dots,n$$

(λ_1, λ_2): are nonnegative regularization parameters.

The first part of the penalty term is the ridge penalty, and the second part is just L_2 (sum of squares) match to the L1 (sum of absolute values) penalty of the fused lasso. The first part shrinks the coefficients toward zero and the second part penalizes roughness. This gives a smooth model [14]. The local linearity penalty is also suggested by (Hawkins & Maboudou-Tchao) in 2013. The local linearity penalty uses the $\|\beta\|_2^2$ and $\sum_{j=2}^{p-1}(\beta_{j+1} - 2\beta_j + \beta_{j-1})^2$ together.

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \underbrace{\lambda_1 \sum_{j=1}^{p}\beta_j^2 + \sum_{j=2}^{p-1}(\beta_{j+1} - 2\beta_j + \beta_{j-1})^2}_{penalty\ term} \qquad (2.\ 20)$$

The first part of the penalty is just the ridge regression, and the second part is to penalize the roughness of the model.

CHAPTER THREE: SRF FOR LINEAR REGRESSION DETALS

In this section the Sparse Ridge Fusion (SRF) will be introduced. First, the coordinate descent algorithm solution is presented to solve the S-Lasso problem. Then it will allow us to a smooth transition to the coordinate descent algorithm solution of the SRF.

The S-Lasso was introduced by Hibiri and van de Geer, and the LARS algorithm was used to solve the problem. The goal here in this paper is to estimate the coefficients β by using the coordinate descent algorithm in lieu of the LARS algorithm.

3. 1. Coordinate Descent algorithm for S-Lasso problem

The model equation with the S-Lasso penalty can be written as follows:

$$f(\beta) = \frac{1}{2} \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=2}^{p}\left(\beta_j - \beta_{j-1}\right)^2 \qquad (3.1)$$

Now the coordinate descent algorithm can use to solve the S-Lasso problem by minimizing the equation (3.1). This gives the following theorem.

Theorem 1:

The coordinate descent solution of the S-Lasso problem is given by

$$For\ j = 1, \hat{\beta}_1 = \frac{S(z_1, \lambda_1)}{1 + \lambda_1} = \begin{cases} \dfrac{z_1 - \lambda_1}{1 + \lambda_2} & , & if\ z_1 > 0\ and\ \lambda_1 < |z_1| \\[2mm] \dfrac{z_1 + \lambda_1}{1 + \lambda_2} & , & if\ z_1 < 0\ and\ \lambda_1 < |z_1| \\[2mm] 0 & , & if \qquad\qquad \lambda_1 \geq |z_1| \end{cases}$$

$$z_1 = \sum_{i=1}^{n} r_i x_{i1} + \lambda_2 \beta_2$$

15

$$For\ j = p, \hat{\beta}_p = \frac{S(z_p, \lambda_1)}{1 + \lambda_1} = \begin{cases} \dfrac{z_p - \lambda_1}{1 + \lambda_2} & , \quad if\ z_p > 0\ and\ \lambda_1 < |z_p| \\ \dfrac{z_p + \lambda_1}{1 + \lambda_2} & , \quad if\ z_p < 0\ and\ \lambda_1 < |z_p| \\ 0 & , \quad if \qquad\quad \lambda_1 \geq |z_p| \end{cases}$$

$$z_p = \sum_{i=1}^{n} r_i x_{ip} + \lambda_2 \beta_{p-1}$$

$$For\ 1 < j < p\ \hat{\beta}_j = \frac{S(z_j, \lambda_1)}{1 + 2\lambda_1} = \begin{cases} \dfrac{z_j - \lambda_1}{1 + 2\lambda_2} & , \quad if\ z_j > 0\ and\ \lambda_1 < |z_j| \\ \dfrac{z_j + \lambda_1}{1 + 2\lambda_2} & , \quad if\ z_j < 0\ and\ \lambda_1 < |z_j| \\ 0 & , \quad if \qquad\quad \lambda_1 \geq |z_j| \end{cases}$$

$$z_j = \sum_{i=1}^{n} r_i x_{ij} + \lambda_2 (\beta_{j+1} + \beta_{j-1})$$

<u>Proof</u>:

The equation (3.1) is equivalent to:

$$f(\beta) = \frac{1}{2} \sum_{i=1}^{n}(r_i - x_{ij}\beta_j)^2 + \lambda_1 \sum_{k \neq j}|\beta_k| + \lambda_1|\beta_j| + \frac{\lambda_2}{2}\sum_{k \neq j}(\beta_k - \beta_{k-1})^2 +$$
$$\frac{\lambda_2}{2}(\beta_{j+1} - \beta_j)^2 + \frac{\lambda_2}{2}(\beta_j - \beta_{j-1})^2 \qquad\qquad (3.2)$$

Where:

$$r_i = y_i - \sum_{k \neq j} x_{ik}\hat{\beta}_k$$

By taking the derivative of the equation (3.2) with respect to β_j,

$$\frac{df(\beta)}{d\ \beta_j} = \begin{cases} -\sum_{i=1}^{n} r_i x_{ij} + \beta_j \sum_{i=1}^{n} x_{ij}^2 + \lambda_1 - \lambda_2(\beta_{j+1} - \beta_j) + \lambda_2(\beta_j - \beta_{j-1}), & if\ \hat{\beta}_j > 0 \\ -\sum_{i=1}^{n} r_i x_{ij} + \beta_j \sum_{i=1}^{n} x_{ij}^2 - \lambda_1 - \lambda_2(\beta_{j+1} - \beta_j) + \lambda_2(\beta_j - \beta_{j-1}), & if\ \hat{\beta}_j < 0 \end{cases}$$

Setting $\frac{df(\beta)}{d\,\beta_j} = 0$ and solving for β_j gives three cases for β_j solutions

$$
\text{For } j = 1, \quad \hat{\beta}_1 =
\begin{cases}
\dfrac{\sum_{i=1}^{n} r_i x_{i1} + \lambda_2\,\beta_2 - \lambda_1}{\sum_{i=1}^{n} x_{i1}^2 + \lambda_2} & , \quad \text{if } \hat{\beta}_1 > 0 \\[4mm]
\dfrac{\sum_{i=1}^{n} r_i x_{i1} + \lambda_2\,\beta_2 + \lambda_1}{\sum_{i=1}^{n} x_{i1}^2 + \lambda_2} & , \quad \text{if } \hat{\beta}_1 < 0
\end{cases}
$$

$$
\text{For } j = p, \quad \hat{\beta}_p =
\begin{cases}
\dfrac{\sum_{i=1}^{n} r_i x_{ip} + \lambda_2\,\beta_{p-1} - \lambda_1}{\sum_{i=1}^{n} x_{ip}^2 + \lambda_2} & , \quad \text{if } \hat{\beta}_p > 0 \\[4mm]
\dfrac{\sum_{i=1}^{n} r_i x_{ip} + \lambda_2\,\beta_{p-1} + \lambda_1}{\sum_{i=1}^{n} x_{ip}^2 + \lambda_2} & , \quad \text{if } \hat{\beta}_p < 0
\end{cases}
$$

$$
\text{For } 1 < j < p, \hat{\beta}_j =
\begin{cases}
\dfrac{\sum_{i=1}^{n} r_i x_{ij} + \lambda_2\,(\beta_{j+1} + \beta_{j-1}) - \lambda_1}{\sum_{i=1}^{n} x_{ij}^2 + 2\lambda_2} & , \quad \text{if } \hat{\beta}_j > 0 \\[4mm]
\dfrac{\sum_{i=1}^{n} r_i x_{ij} + \lambda_2\,(\beta_{j+1} + \beta_{j-1}) + \lambda_1}{\sum_{i=1}^{n} x_{ij}^2 + 2\lambda_2} & , \quad \text{if } \hat{\beta}_j < 0
\end{cases}
$$

As $\sum_{i=1}^{n} x_{ij}^2 = 1 \ \ for \ j = \{1, 2, \dots, p\}$, the above cases simplify to

$$
\text{For } j = 1, \quad \hat{\beta}_1 =
\begin{cases}
\dfrac{\sum_{i=1}^{n} r_i x_{i1} + \lambda_2\,\beta_2 - \lambda_1}{1 + \lambda_2} & , \quad \text{if } \hat{\beta}_1 > 0 \\[4mm]
\dfrac{\sum_{i=1}^{n} r_i x_{i1} + \lambda_2\,\beta_2 + \lambda_1}{1 + \lambda_2} & , \quad \text{if } \hat{\beta}_1 < 0
\end{cases}
$$

$$
\text{For } j = p, \quad \hat{\beta}_p =
\begin{cases}
\dfrac{\sum_{i=1}^{n} r_i x_{ip} + \lambda_2\,\beta_{p-1} - \lambda_1}{1 + \lambda_2} & , \quad \text{if } \hat{\beta}_p > 0 \\[4mm]
\dfrac{\sum_{i=1}^{n} r_i x_{ip} + \lambda_2\,\beta_{p-1} + \lambda_1}{1 + \lambda_2} & , \quad \text{if } \hat{\beta}_p < 0
\end{cases}
$$

$$
\text{For } j = 1, \quad \hat{\beta}_j =
\begin{cases}
\dfrac{\sum_{i=1}^{n} r_i x_{ij} + \lambda_2\,(\beta_{j+1} + \beta_{j-1}) - \lambda_1}{1 + 2\lambda_2} & , \quad \text{if } \hat{\beta}_j > 0 \\[4mm]
\dfrac{\sum_{i=1}^{n} r_i x_{ij} + \lambda_2\,(\beta_{j+1} + \beta_{j-1}) + \lambda_1}{1 + 2\lambda_2} & , \quad \text{if } \hat{\beta}_j < 0
\end{cases}
$$

By using the soft threshold operator, the solution is given by:

$$\text{For } j = 1, \quad \hat{\beta}_1 = \frac{S(z_1, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_1 - \lambda_1}{1 + \lambda_2} & , \quad if \ z_1 > 0 \ and \ \lambda_1 < |z_1| \\ \dfrac{z_1 + \lambda_1}{1 + \lambda_2} & , \quad if \ z_1 < 0 \ and \ \lambda_1 < |z_1| \\ 0 & , \quad if \qquad \lambda_1 \geq |z_1| \end{cases}$$

Where: $z_1 = \sum_{i=1}^{n} r_i x_{i1} + \lambda_2 \beta_2$

$$\text{For } j = p, \quad \hat{\beta}_p = \frac{S(z_p, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_p - \lambda_1}{1 + \lambda_2} & , \quad if \ z_p > 0 \ and \ \lambda_1 < |z_p| \\ \dfrac{z_p + \lambda_1}{1 + \lambda_2} & , \quad if \ z_p < 0 \ and \ \lambda_1 < |z_p| \\ 0 & , \quad if \qquad \lambda_1 \geq |z_p| \end{cases}$$

Where: $z_p = \sum_{i=1}^{n} r_i x_{ip} + \lambda_2 \beta_{p-1}$

$$\text{For } 1 < j < p, \quad \hat{\beta}_j = \frac{S(z_j, \lambda_1)}{1 + 2\lambda_2} = \begin{cases} \dfrac{z_j - \lambda_1}{1 + 2\lambda_2} & , \quad if \ z_j > 0 \ and \ \lambda_1 < |z_j| \\ \dfrac{z_j + \lambda_1}{1 + 2\lambda_2} & , \quad if \ z_j < 0 \ and \ \lambda_1 < |z_j| \\ 0 & , \quad if \qquad \lambda_1 \geq |z_j| \end{cases}$$

Where: $z_j = \sum_{i=1}^{n} r_i x_{ij} + \lambda_2 \left(\beta_{j+1} + \beta_{j-1} \right)$

3.2. Sparse Ridge Fusion (SRF) Penalty

In this section, a new proposal (SRF) is introduced, and the coordinate descent algorithm is used to solve the problem.

The penalty of $\sum_{j=2}^{p-1} \left(\beta_{j+1} - 2\beta_j + \beta_{j-1} \right)^2$ applied to the regression problems was first used in [14] as a local linearity penalty which was combined with the L_2- ridge penalty. The SRF penalty is $\lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{2} \sum_{j=2}^{p-1} \left(\beta_{j+1} - 2\beta_j + \beta_{j-1} \right)^2$. So, the interest is to minimize

$$\hat{\beta}_{SRF} = arg\ min\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \frac{\lambda_2}{2}\sum_{j=2}^{p-1}(\beta_{j+1} - 2\beta_j + \beta_{j-1})^2\right\} \quad (3.3)$$

Where:

$$i = 1, 2, \ldots, n \quad , \quad j = 1, 2, \ldots, p$$

$\sum_{j=1}^{p}|\beta_j|$: is the L$_1$-Lasso penalty which is used for sparsity in the coefficients. It shrinks

the fitted coefficients towards zero. The second part of the penalty $\sum_{j=2}^{p-1}(\beta_{j+1} - 2\beta_j +$

$\beta_{j-1})^2$ penalizes roughness.

3.2.1 The Sparse Ridge Fusion Estimate

Given data set (X, y)

Where:

$X_{(nxp)}$ is a design matrix, $\quad X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}_{nxp}$

$y_{(nx1)}$ is the response variable, $\quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{nx1}$

(λ_1, λ_2): are non-negative values

Then the residual sum square function with SRF penalty is given as bellow:

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \frac{\lambda_2}{2}\sum_{j=2}^{p-1}(\beta_{j+1} - 2\beta_j + \beta_{j-1})^2 \quad (3.4)$$

Then, using the coordinate descent algorithm to solve the SRF problem by

minimizing the equation (3.4) gives the following theorem.

Theorem 2:

The coordinate descent solution of the SRF is given by following result:

$$\text{For } j = 1, \quad \hat{\beta}_1 = \frac{S(z_1, \ \lambda_1)}{1 + \lambda_1} = \begin{cases} \dfrac{z_1 - \lambda_1}{1 + \lambda_2} & , \quad if \quad z_1 > 0 , \lambda_1 < |z_1| \\[3mm] \dfrac{z_1 + \lambda_1}{1 + \lambda_2} & , \quad if \quad z_1 < 0 , \lambda_1 < |z_1| \\[3mm] 0 & , \quad if \quad\quad\quad \lambda_1 \geq |z_1| \end{cases}$$

$$z_1 = \sum_{i=1}^{n} r_i x_{i1} + 2\lambda_2 \, \beta_2 - \lambda_2 \beta_3$$

$$\text{For } j = p, \quad \hat{\beta}_p = \frac{S(z_p, \ \lambda_1)}{1 + \lambda_1} = \begin{cases} \dfrac{z_p - \lambda_1}{1 + \lambda_2} & , \quad if \quad z_p > 0 , \lambda_1 < \left|z_p\right| \\[3mm] \dfrac{z_p + \lambda_1}{1 + \lambda_2} & , \quad if \quad z_p < 0 , \lambda_1 < \left|z_p\right| \\[3mm] 0 & , \quad if \quad\quad\quad \lambda_1 \geq \left|z_p\right| \end{cases}$$

$$z_p = \sum_{i=1}^{n} r_i x_{ip} + 2\lambda_2 \, \beta_{p-1} - \lambda_2 \beta_{p-2}$$

$$\text{For } j = 2, \quad \hat{\beta}_2 = \frac{S(z_2, \ \lambda_1)}{1 + 5\lambda_1} = \begin{cases} \dfrac{z_2 - \lambda_1}{1 + 5\lambda_2} & , \quad if \quad z_2 > 0 , \lambda_1 < |z_2| \\[3mm] \dfrac{z_2 + \lambda_1}{1 + 5\lambda_2} & , \quad if \quad z_2 < 0 , \lambda_1 < |z_2| \\[3mm] 0 & , \quad if \quad\quad\quad \lambda_1 \geq |z_2| \end{cases}$$

$$z_2 = \sum_{i=1}^{n} r_i \, x_{i2} + 4\lambda_2 \, \beta_3 - 2\lambda_2 \, \beta_1 - \beta_4$$

$$\text{For } j = p - 1, \quad \hat{\beta}_{p-1} = \frac{S(z_{p-1}, \ \lambda_1)}{1 + 5\lambda_1} = \begin{cases} \dfrac{z_{p-1} - \lambda_1}{1 + 5\lambda_2} & , if \quad z_{p-1} > 0 , \lambda_1 < \left|z_{p-1}\right| \\[3mm] \dfrac{z_{p-1} + \lambda_1}{1 + 5\lambda_2} & , if \quad z_{p-1} < 0 , \lambda_1 < \left|z_{p-1}\right| \\[3mm] 0 & , if \quad\quad\quad \lambda_1 \geq \left|z_{p-1}\right| \end{cases}$$

$$z_{p-1} = \sum_{i=1}^{n} r_i \, x_{ip-1} + 4\lambda_2 \, \beta_{p-2} - 2\lambda_2 \, \beta_p - \beta_{p-3}$$

$$For\ 3 \le j \le p - 2, \ \hat{\beta}_j = \frac{S(z_j, \ \lambda_1)}{1 + 5\lambda_1} = \begin{cases} \dfrac{z_j - \lambda_1}{1 + 6\lambda_2} & , \ if \quad z_j > 0, \lambda_1 < |z_j| \\ \dfrac{z_j + \lambda_1}{1 + 6\lambda_2} & , \ if \quad z_j < 0, \lambda_1 < |z_j| \\ 0 & , \ if \qquad\quad \lambda_1 \ge |z_j| \end{cases}$$

$$z_j = \sum_{i=1}^{n} r_i x_{ij} + 4\lambda_2(\beta_{j+1} + \beta_{j-1}) - \lambda_2(\beta_{j+2} + \beta_{j-2}) \quad for\ 3 \le j \le p - 2$$

Proof:

The equation (3.4) can be written as following equation:

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}(r_i - \beta_j x_{ij})^2 + \lambda_1 \sum_{k \ne j}|\beta_k| + \lambda_1|\beta_j| + \frac{\lambda_2}{2}\sum_{k \ne j}(\beta_{k+1} - 2\beta_k + \beta_{k-1})^2$$

$$+ \frac{\lambda_2}{2}(\beta_{j+1} - 2\beta_j + \beta_{j-1})^2 + \frac{\lambda_2}{2}(\beta_j - 2\beta_{j-1} + \beta_{j-2})^2 + \frac{\lambda_2}{2}(\beta_{j+2} - 2\beta_{j+1} + \beta_j)^2 \qquad (3.5)$$

The derivative of equation (3.5) with respect to β_j gives

$$\frac{df(\beta)}{d\ \beta_j}$$

$$= \begin{cases} -\sum_{i=1}^{n} r_i x_{ij} + \beta_j \sum_{i=1}^{n} x_{ij}^2 + \lambda_1 - 2\lambda_2(\beta_{j+1} - 2\beta_j + \beta_{j-1}) + \lambda_2(\beta_j - 2\beta_{j-1} + \beta_{j-2}) + \lambda_2(\beta_{j+2} - 2\beta_{j+1} + \beta_j), \beta_j > 0 \\ -\sum_{i=1}^{n} r_i x_{ij} + \beta_j \sum_{i=1}^{n} x_{ij}^2 - \lambda_1 - 2\lambda_2(\beta_{j+1} - 2\beta_j + \beta_{j-1}) + \lambda_2(\beta_j - 2\beta_{j-1} + \beta_{j-2}) + \lambda_2(\beta_{j+2} - 2\beta_{j+1} + \beta_j), \beta_j < 0 \end{cases}$$

Setting $\frac{df(\beta)}{d\ \beta_j} = 0$ and solving β_j gives five cases for β_j solutions

$$For\ j = 1, \hat{\beta}_1 = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i x_{i1} - \lambda_1 + 2\lambda_2\beta_2 - \lambda_2\beta_3}{\sum_{i=1}^{n} x_{i1}^2 + \lambda_2} & , \ if \ \hat{\beta}_1 < 0 \\ \dfrac{\sum_{i=1}^{n} r_i x_{i1} + \lambda_1 + 2\lambda_2\beta_2 - \lambda_2\beta_3}{\sum_{i=1}^{n} x_{i1}^2 + \lambda_2} & , \quad if \ \hat{\beta}_1 < 0 \end{cases}$$

21

$$For\ j = p, \hat{\beta}_p = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i\, x_{ip} - \lambda_1 + 2\lambda_2\beta_{p-1} - \lambda_2\beta_{p-2}}{\sum_{i=1}^{n} x_{ip}^2 + \lambda_2} & ,\quad if\ \ \hat{\beta}_p < 0 \\[4mm] \dfrac{\sum_{i=1}^{n} r_i\, x_{ip} + \lambda_1 + 2\lambda_2\beta_{p-1} - \lambda_2\beta_{p-2}}{\sum_{i=1}^{n} x_{ip}^2 + \lambda_2} & ,\quad if\ \ \hat{\beta}_p < 0 \end{cases}$$

$$For\ j = 2, \hat{\beta}_2 = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i\, x_{i2} - \lambda_1 + 4\lambda_2\, \beta_3 - 2\lambda_2\, \beta_1 - \beta_4}{\sum_{i=1}^{n} x_{i2}^2 + 5\,\lambda_2} & ,\quad if\ \ \hat{\beta}_2 < 0 \\[4mm] \dfrac{\sum_{i=1}^{n} r_i\, x_{i2} + \lambda_1 + 4\lambda_2\, \beta_3 - 2\lambda_2\, \beta_1 - \beta_4}{\sum_{i=1}^{n} x_{i2}^2 + 5\,\lambda_2} & ,\quad if\ \ \hat{\beta}_2 < 0 \end{cases}$$

$$For\ j = p - 1, \hat{\beta}_{p-1} = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i\, x_{ip-1} - \lambda_1 + 4\lambda_2\, \beta_{p-2} - 2\lambda_2\, \beta_p - \beta_{p-3}}{\sum_{i=1}^{n} x_{ip-1}^2 + 5\,\lambda_2} & if\ \ \hat{\beta}_{p-1} < 0 \\[4mm] \dfrac{\sum_{i=1}^{n} r_i\, x_{ip-1} + \lambda_1 + 4\lambda_2\, \beta_{p-2} - 2\lambda_2\, \beta_p - \beta_{p-3}}{\sum_{i=1}^{n} x_{ip-1}^2 + 5\,\lambda_2} & if\ \ \hat{\beta}_{p-1} < 0 \end{cases}$$

For $3 \le j \le p - 2$

$$\hat{\beta}_j = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i\, x_{ij} - \lambda_1 + 4\lambda_2\big(\beta_{j+1} + \beta_{j-1}\big) - \lambda_2\big(\beta_{j+2} + \beta_{j-2}\big)}{\sum_{i=1}^{n} x_{ij}^2 + 6\,\lambda_2} & ,\quad if\ \hat{\beta}_j < 0 \\[4mm] \dfrac{\sum_{i=1}^{n} r_i\, x_{ij} + \lambda_1 + 4\lambda_2\big(\beta_{j+1} + \beta_{j-1}\big) - \lambda_2\big(\beta_{j+2} + \beta_{j-2}\big)}{\sum_{i=1}^{n} x_{ij}^2 + 6\,\lambda_2} & ,\quad if\ \hat{\beta}_j < 0 \end{cases}$$

As $\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = \{1, 2, \ldots, p\}$ the above cases simplify to

$$For\ j = 1, \hat{\beta}_1 = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i\, x_{i1} - \lambda_1 + 2\lambda_2\beta_2 - \lambda_2\beta_3}{1 + \lambda_2} & ,\quad if\ \ \hat{\beta}_1 < 0 \\[4mm] \dfrac{\sum_{i=1}^{n} r_i\, x_{i1} + \lambda_1 + 2\lambda_2\beta_2 - \lambda_2\beta_3}{1 + \lambda_2} & ,\quad if\ \ \hat{\beta}_1 < 0 \end{cases}$$

$$For\ j = p, \hat{\beta}_p = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i\, x_{ip} - \lambda_1 + 2\lambda_2\beta_{p-1} - \lambda_2\beta_{p-2}}{1 + \lambda_2} & ,\quad if\ \ \hat{\beta}_p < 0 \\[4mm] \dfrac{\sum_{i=1}^{n} r_i\, x_{ip} + \lambda_1 + 2\lambda_2\beta_{p-1} - \lambda_2\beta_{p-2}}{1 + \lambda_2} & ,\quad if\ \ \hat{\beta}_p < 0 \end{cases}$$

$$\text{For } j = 2, \qquad \hat{\beta}_2 = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i\, x_{i2} - \lambda_1 + 4\lambda_2\, \beta_3 - 2\lambda_2\, \beta_1 - \beta_4}{1 + 5\,\lambda_2} & , \ if \quad \hat{\beta}_2 > 0 \\[2mm] \dfrac{\sum_{i=1}^{n} r_i\, x_{i2} + \lambda_1 + 4\lambda_2\, \beta_3 - 2\lambda_2\, \beta_1 - \beta_4}{1 + 5\,\lambda_2} & , \ if \quad \hat{\beta}_2 < 0 \end{cases}$$

$$\text{For } j = p - 1, \hat{\beta}_{p-1} = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i\, x_{ip-1} - \lambda_1 + 4\lambda_2\, \beta_{p-2} - 2\lambda_2\, \beta_p - \beta_{p-3}}{1 + 5\,\lambda_2} & if \ \hat{\beta}_{p-1} > 0 \\[2mm] \dfrac{\sum_{i=1}^{n} r_i\, x_{ip-1} + \lambda_1 + 4\lambda_2\, \beta_{p-2} - 2\lambda_2\, \beta_p - \beta_{p-3}}{1 + 5\,\lambda_2} & if \ \hat{\beta}_{p-1} < 0 \end{cases}$$

For $3 \leq j \leq p - 2$

$$\hat{\beta}_j = \begin{cases} \dfrac{\sum_{i=1}^{n} r_i\, x_{ij} - \lambda_1 + 4\lambda_2\big(\beta_{j+1} + \beta_{j-1}\big) - \lambda_2\big(\beta_{j+2} + \beta_{j-2}\big)}{1 + 6\,\lambda_2} & , \ if \ \hat{\beta}_j > 0 \\[2mm] \dfrac{\sum_{i=1}^{n} r_i\, x_{ij} + \lambda_1 + 4\lambda_2\big(\beta_{j+1} + \beta_{j-1}\big) - \lambda_2\big(\beta_{j+2} + \beta_{j-2}\big)}{1 + 6\,\lambda_2} & , \ if \ \hat{\beta}_j < 0 \end{cases}$$

By using the soft threshold operator, the solution is given by

$$\text{For } j = 1, \qquad \hat{\beta}_1 = \frac{S(z_1,\ \lambda_1)}{1 + \lambda_1} = \begin{cases} \dfrac{z_1 - \lambda_1}{1 + \lambda_2} & , \quad if \quad z_1 > 0\,, \lambda_1 < |z_1| \\[2mm] \dfrac{z_1 + \lambda_1}{1 + \lambda_2} & , \quad if \quad z_1 < 0\,, \lambda_1 < |z_1| \\[2mm] 0 & , \quad if \qquad \lambda_1 \geq |z_1| \end{cases}$$

$$z_1 = \sum_{i=1}^{n} r_i x_{i1} + 2\lambda_2\, \beta_2 - \lambda_2 \beta_3$$

$$\text{For } j = p, \qquad \hat{\beta}_p = \frac{S(z_p,\ \lambda_1)}{1 + \lambda_1} = \begin{cases} \dfrac{z_p - \lambda_1}{1 + \lambda_2} & , \quad if \quad z_p > 0\,, \lambda_1 < |z_p| \\[2mm] \dfrac{z_p + \lambda_1}{1 + \lambda_2} & , \quad if \quad z_p < 0\,, \lambda_1 < |z_p| \\[2mm] 0 & , \quad if \qquad \lambda_1 \geq |z_p| \end{cases}$$

$$z_p = \sum_{i=1}^{n} r_i x_{ip} + 2\lambda_2\, \beta_{p-1} - \lambda_2 \beta_{p-2}$$

$$\text{For } j = 2, \quad \hat{\beta}_2 = \frac{S(z_2, \lambda_1)}{1 + 5\lambda_2} = \begin{cases} \dfrac{z_2 - \lambda_1}{1 + 5\lambda_2} & , \ if \ z_2 > 0 \ and \ \lambda_1 < |z_2| \\[2mm] \dfrac{z_2 + \lambda_1}{1 + 5\lambda_2} & , \ if \ z_2 < 0 \ and \ \lambda_1 < |z_2| \\[2mm] 0 & , \ if \quad\quad\quad \lambda_1 \geq |z_2| \end{cases}$$

$$z_2 = \sum_{i=1}^{n} r_i \, x_{i2} + 4\lambda_2 \, \beta_3 - 2\lambda_2 \, \beta_1 - \beta_4$$

$$\text{For } j = p - 1, \quad \hat{\beta}_{p-1} = \frac{S(z_{p-1}, \lambda_1)}{1 + 5\lambda_2} = \begin{cases} \dfrac{z_{p-1} - \lambda_1}{1 + 5\lambda_2} & , \ if \ z_{p-1} > 0 , \lambda_1 < |z_{p-1}| \\[2mm] \dfrac{z_{p-1} + \lambda_1}{1 + 5\lambda_2} & , \ if \ z_{p-1} < 0 , \lambda_1 < |z_{p-1}| \\[2mm] 0 & , \ if \quad\quad\quad \lambda_1 \geq |z_{p-1}| \end{cases}$$

$$z_{p-1} = \sum_{i=1}^{n} r_i \, x_{ip-1} + 4\lambda_2 \, \beta_{p-2} - 2\lambda_2 \, \beta_p - \beta_{p-3}$$

$$\text{For } 3 \leq j \leq p - 2, \quad \hat{\beta}_j = \frac{S(z_j, \lambda_1)}{1 + 6\lambda_2} = \begin{cases} \dfrac{z_j - \lambda_1}{1 + 6\lambda_2} & , \ if \ z_j > 0 , \lambda_1 < |z_j| \\[2mm] \dfrac{z_j + \lambda_1}{1 + 6\lambda_2} & , \ if \ z_j < 0 , \lambda_1 < |z_j| \\[2mm] 0 & , \ if \quad\quad\quad \lambda_1 \geq |z_j| \end{cases}$$

$$z_j = \sum_{i=1}^{n} r_i \, x_{ij} + 4\lambda_2 \big(\beta_{j+1} + \beta_{j-1}\big) - \lambda_2 \big(\beta_{j+2} + \beta_{j-2}\big) \quad for \ 3 \leq j \leq p - 2$$

CHAPTER FOUR: COMPUTATIONS

In this section, an augmented data set is created in order to streamline computation of the coordinate descent algorithm solution to solve the S-Lasso and SRF problems.

4.1. Augmented data set for S-Lasso

Given data set (X, y)

Where:

X: is a design matrix n x p

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}_{nxp}$$

y: is an n observations vector

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_3 \end{pmatrix}_{nx1}$$

And $(\lambda_1, \lambda_2) \geq 0$

Now define

$$X^* = \begin{pmatrix} X \\ J \end{pmatrix} \quad , \text{and} \quad y^* = \begin{pmatrix} y \\ 0 \end{pmatrix}$$

Where:

0 is a vector of size p which contains only zeros and J is the p $\times$ p matrix defined as

$$J = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ \sqrt{\lambda_2} & -\sqrt{\lambda_2} & 0 & 0 & \ddots & 0 \\ 0 & \sqrt{\lambda_2} & -\sqrt{\lambda_2} & 0 & \ddots & 0 \\ 0 & 0 & \sqrt{\lambda_2} & -\sqrt{\lambda_2} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & \sqrt{\lambda_2} & -\sqrt{\lambda_2} \end{bmatrix}_{pxp}$$

X^*: becomes a $(n + p) \times p$ design matrix, and y^* becomes $(n + p) \times 1$ vector of observations.

According to (X^*, y^*) data set, the lasso penalty will be:

$$f(\beta) = \frac{1}{2} \sum_{i=1}^{n^*} \left(y_i^* - \sum_{j=1}^{p} x_{ij}^* \beta_j^* \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j^*| \qquad (4.1)$$

The minimizing of the equation (4.1) brings the following theorem.

Theorem 3:

The S-Lasso solves the lasso type problem by using an augmented data set artificially to define a new data set (X^*, y^*), and the solutions are:

$$For\ j = 1, \hat{\beta}_1^* = \frac{S(z_1, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_1 - \lambda_1}{1 + \lambda_2} & , & if & z_1 > 0, \lambda_1 < |z_1| \\[2mm] \dfrac{z_1 + \lambda_1}{1 + \lambda_2} & , & if & z_1 < 0, \lambda_1 < |z_1| \\[2mm] 0 & , & if & \lambda_1 \geq |z_1| \end{cases}$$

$$z_1 = \sum_{i=1}^{n} r_i x_{i1} + \lambda_2 \beta_2^*$$

$$For\ j = p, \hat{\beta}_p^* = \frac{S(z_p, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_p - \lambda_1}{1 + \lambda_2} & , & if & z_p > 0, \lambda_1 < |z_p| \\[2mm] \dfrac{z_p + \lambda_1}{1 + \lambda_2} & , & if & z_p < 0, \lambda_1 < |z_p| \\[2mm] 0 & , & if & \lambda_1 \geq |z_p| \end{cases}$$

$$z_p = \sum_{i=1}^{n} r_i x_{ip} + \lambda_2 \beta_{p-1}^*$$

$$For\ 1 < j < p, \hat{\beta}_j^* = \frac{S(z_j, \lambda_1)}{1 + 2\lambda_2} = \begin{cases} \dfrac{z_j - \lambda_1}{1 + 2\lambda_2} & , & if & z_j > 0, \lambda_1 < |z_j| \\[2mm] \dfrac{z_j + \lambda_1}{1 + 2\lambda_2} & , & if & z_j < 0, \lambda_1 < |z_j| \\[2mm] 0 & , & if & \lambda_1 \geq |z_j| \end{cases}$$

$$z_j = \sum_{i=1}^{n} r_i x_{ij} + \lambda_2 \left(\beta_{j+1}^* - \beta_{j-1}^*\right)$$

<u>Proof</u>:

The equation 4.1 is equivalent to the below:

$$f(\beta) = \frac{1}{2} \sum_{i=1}^{n^*} \left(r_i^* - x_{ij}^* \beta_j^*\right)^2 + \lambda_1 \sum_{k \neq j} |\beta_k^*| + \lambda_1 |\beta_j^*| \qquad (4.2)$$

Where:

$$r_i^* = y_i^* - \sum_{k \neq j} x_{ik}^* \hat{\beta}_k^*$$

$$n^* = n + p$$

While the sample size X^* is $n + p$, The S-Lasso can possibly select all p predictor variables in all situations.

The derivative of equation (4.2) with respect to β_j^* gives

$$\frac{df(x)}{d\,\beta_j^*} = \begin{cases} -\sum_{i=1}^{n^*} r_i^* x_{ij}^* + \beta_j^* \sum_{i=1}^{n^*} x_{ij}^{*\,2} + \lambda_1 & , & if & |\beta^*| > 0 \\ -\sum_{i=1}^{n^*} r_i^* x_{ij}^* + \beta_j^* \sum_{i=1}^{n^*} x_{ij}^{*\,2} + \lambda_1 & , & if & |\beta^*| < 0 \end{cases} \qquad (4.3)$$

In the equation (4.3)

$$\sum_{i=1}^{n^*} r_i^* x_{ij}^* = \sum_{i=1}^{n} r_i x_{ij} + \sum_{i=1}^{p} r_i J_{ij} \qquad (4.4)$$

$$\sum_{i=1}^{n^*} x_{ij}^{*\,2} = \sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{p} J_{ij}^2 \qquad (4.5)$$

By plugging the equations (4.4) and (4.5) into (4.3), the derivatives become as below:

$$\frac{df(x)}{d\,\beta_j^*} = \begin{cases} -\left(\sum_{i=1}^{n} r_i x_{ij} + \sum_{i=1}^{p} r_i J_{ij}\right) + \beta_j^* \left(\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{p} J_{ij}^2\right) + \lambda_1 & if & |\beta^*| > 0 \\ -\left(\sum_{i=1}^{n} r_i x_{ij} + \sum_{i=1}^{p} r_i J_{ij}\right) + \beta_j^* \left(\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{p} J_{ij}^2\right) - \lambda_1 & if & |\beta^*| < 0 \end{cases}$$

$$(4.6)$$

Setting $\frac{df(x)}{d\,\beta_j^*} = 0$ and solving β_j^* gives three cases for β_j^* solutions

$$\text{For } j = 1, \quad \hat{\beta}_1^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{i1} + (\sqrt{\lambda_2}\,\beta_2^*)\sqrt{\lambda_2} - \lambda_1}{\sum_{i=1}^n x_{i1}^2 + (\sqrt{\lambda_2})^2} & , \quad if \quad \hat{\beta}_1^* > 0 \\[3mm] \dfrac{\sum_{i=1}^n r_i x_{i1} + (\sqrt{\lambda_2}\,\beta_2^*)\sqrt{\lambda_2} + \lambda_1}{\sum_{i=1}^n x_{i1}^2 + (\sqrt{\lambda_2})^2} & , \quad if \quad \hat{\beta}_1^* < 0 \end{cases}$$

$$\text{For } j = p, \quad \hat{\beta}_p^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ip} + (-\sqrt{\lambda_2}\,\beta_{p-1}^*)(-\sqrt{\lambda_2}) - \lambda_1}{\sum_{i=1}^n x_{ip}^2 + (-\sqrt{\lambda_2})^2} & , \quad if \quad \hat{\beta}_p^* > 0 \\[3mm] \dfrac{\sum_{i=1}^n r_i x_{ip} + (-\sqrt{\lambda_2}\,\beta_{p-1}^*)(-\sqrt{\lambda_2}) + \lambda_1}{\sum_{i=1}^n x_{ip}^2 + (-\sqrt{\lambda_2})^2} & , \quad if \quad \hat{\beta}_p^* < 0 \end{cases}$$

$$\text{For } 1 < j < p, \hat{\beta}_j^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ij} + (-\sqrt{\lambda_2}\,\beta_{j-1}^*)(-\sqrt{\lambda_2}) + (\sqrt{\lambda_2}\beta_{j+1}^*) - \lambda_1}{\sum_{i=1}^n x_{ij}^2 + (-\sqrt{\lambda_2})^2 + (\sqrt{\lambda_2})^2} & if \; \hat{\beta}_j^* > 0 \\[3mm] \dfrac{\sum_{i=1}^n r_i x_{ij} + (-\sqrt{\lambda_2}\,\beta_{j-1}^*)(-\sqrt{\lambda_2}) + (\sqrt{\lambda_2}\beta_{j+1}^*) + \lambda_1}{\sum_{i=1}^n x_{ij}^2 + (-\sqrt{\lambda_2})^2 + (\sqrt{\lambda_2})^2} & if \; \hat{\beta}_j^* < 0 \end{cases}$$

As $\sum_{i=1}^n x_{ij}^2 = 1 \; for \; j = 1, 2, \dots, p$ the above cases simplify to:

$$\text{For } j = 1, \quad \hat{\beta}_1^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{i1} + \lambda_2\,\beta_2^* - \lambda_1}{1 + \lambda_2} & , \quad if \quad \hat{\beta}_1^* > 0 \\[3mm] \dfrac{\sum_{i=1}^n r_i x_{i1} + \lambda_2\,\beta_2^* + \lambda_1}{1 + \lambda_2} & , \quad if \quad \hat{\beta}_1^* < 0 \end{cases}$$

$$\text{For } j = p, \quad \hat{\beta}_p^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ip} + \lambda_2\,\beta_{p-1}^* - \lambda_1}{1 + \lambda_2} & , \quad if \quad \hat{\beta}_p^* > 0 \\[3mm] \dfrac{\sum_{i=1}^n r_i x_{ip} + \lambda_2\,\beta_{p-1}^* + \lambda_1}{1 + \lambda_2} & , \quad if \quad \hat{\beta}_p^* < 0 \end{cases}$$

$$\text{For } 1 < j < p, \hat{\beta}_j^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ij} + \lambda_2(\beta_{j-1}^* + \beta_{j+1}^*) - \lambda_1}{1 + 2\lambda_2} & , \quad if \quad \hat{\beta}_j^* > 0 \\[3mm] \dfrac{\sum_{i=1}^n r_i x_{ij} + \lambda_2(\beta_{j-1}^* + \beta_{j+1}^*) - \lambda_1}{1 + 2\lambda_2} & , \quad if \quad \hat{\beta}_j^* < 0 \end{cases}$$

By using the soft threshold operator, the solution is given by

$$For\ j = 1,\ \hat{\beta}_1^* = \frac{S(z_1, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_1 - \lambda_1}{1 + \lambda_2} & , \quad if\ z_1 > 0\ and\ \lambda_1 < |z_1| \\ \dfrac{z_1 + \lambda_1}{1 + \lambda_2} & , \quad if\ z_1 < 0\ and\ \lambda_1 < |z_1| \\ 0 & , \quad if \qquad\qquad \lambda_1 \geq |z_1| \end{cases}$$

$$z_1 = \sum_{i=1}^{n} r_i x_{i1} + \lambda_2\, \beta_2^*$$

$$For\ j = p, \hat{\beta}_p^* = \frac{S(z_p, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_p - \lambda_1}{1 + \lambda_2} & , \quad if\ z_p > 0\ and\ \lambda_1 < |z_p| \\ \dfrac{z_p + \lambda_1}{1 + \lambda_2} & , \quad if\ z_p < 0\ and\ \lambda_1 < |z_p| \\ 0 & , \quad if \qquad\qquad \lambda_1 \geq |\hat{\beta}_p^*| \end{cases}$$

$$z_p = \sum_{i=1}^{n} r_i x_{ip} + \lambda_2\, \beta_{p-1}^*$$

$$For\ 1 < j < p, \hat{\beta}_j^* = \frac{S(z_j, \lambda_1)}{1 + 2\,\lambda_2} = \begin{cases} \dfrac{z_j - \lambda_1}{1 + 2\,\lambda_2} & , \quad if\ z_j > 0\ and\ \lambda_1 < |z_j| \\ \dfrac{z_j + \lambda_1}{1 + 2\,\lambda_2} & , \quad if\ z_j < 0\ and\ \lambda_1 < |z_j| \\ 0 & , \quad if \qquad\qquad \lambda_1 \geq |z_j| \end{cases}$$

$$z_j = \sum_{i=1}^{n} r_i x_{ij} + \lambda_2\, (\beta_{j+1}^* - \beta_{j-1}^*)$$

4.2. Augmented Data Set Estimate for SRF problem

Identical to the previous method used to solve the S-Lasso problem, an augmented data is created here to simplify computation of the coordinate descent solution to solve the SRF problem.

Therefore,

$$X^* = \begin{pmatrix} X \\ K \end{pmatrix}$$

$$y^* = \begin{pmatrix} y \\ 0 \end{pmatrix}$$

Where:

0 is a vector of size p which contains only zeros and K is the $p \times p$ matrix

$$K = \begin{bmatrix} 0 & 0 & \cdot & \cdot & \cdot & & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & & 0 \\ -\sqrt{\lambda_2} & 2\sqrt{\lambda_2} & \sqrt{\lambda_2} & 0 & \cdots & & \cdot \\ 0 & -\sqrt{\lambda_2} & 2\sqrt{\lambda_2} & \sqrt{\lambda_2} & 0.. & & \cdot \\ \vdots & \cdots & \ddots & \ddots & \ddots & & \cdot \\ 0 & \cdots & 0 & -\sqrt{\lambda_2} & 2\sqrt{\lambda_2} & -\sqrt{\lambda_2} \end{bmatrix}_{pxp}$$

Thus, X^* becomes a $n^* \times p$ design matrix, and y^* becomes an n^* x 1 vector of observations, and

$$n^* = n + p$$

Since the sample size in the augmented problem is $n + p$, the sparse ridge fusion (SRF) can possibly select all p predictors in all situations.

According to (X^*, y^*) data set, the lasso penalty will be

$$f(\beta^*) = \frac{1}{2} \sum_{i=1}^{n^*} \left(y_i^* - \sum_{j=1}^{p} x_{ij}^* \beta_j^*\right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j^*| \tag{4.7}$$

Where:

$$i = 1, 2, \ldots, n^* \qquad and \qquad j = 1, 2, \ldots, p$$

The minimizing of the equation (4.7) brings the following theorem.

Theorem 4:

The SRF solves the lasso type problem by using an augmented data set (X^*, y^*), and the solutions are:

$$\text{For } j = 1, \quad \hat{\beta}_1 = \frac{S(z_1, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_1 - \lambda_1}{1 + \lambda_2} & , \quad if \quad z_1 > 0 \, , \lambda_1 < |z_1| \\[2mm] \dfrac{z_1 + \lambda_1}{1 + \lambda_2} & , \quad if \quad z_1 < 0 \, , \lambda_1 < |z_1| \\[2mm] 0 & , \quad if \quad \qquad \lambda_1 \geq |z_1| \end{cases}$$

$$z_1 = \sum_{i=1}^{n} r_i x_{i1} + 2\lambda_2 \, \beta_2 - \lambda_2 \beta_3$$

$$\text{For } j = p, \quad \hat{\beta}_p = \frac{S(z_1, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_p - \lambda_1}{1 + \lambda_2} & , \quad if \quad z_p > 0 \, , \lambda_1 < |z_p| \\[2mm] \dfrac{z_p + \lambda_1}{1 + \lambda_2} & , \quad if \quad z_p < 0 \, , \lambda_1 < |z_p| \\[2mm] 0 & , \quad if \quad \qquad \lambda_1 \geq |z_p| \end{cases}$$

$$z_p = \sum_{i=1}^{n} r_i x_{ip} + 2\lambda_2 \, \beta_{p-1} - \lambda_2 \beta_{p-2}$$

$$\text{For } j = 2, \quad \hat{\beta}_2 = \frac{S(z_2, \lambda_1)}{1 + 5\lambda_2} = \begin{cases} \dfrac{z_2 - \lambda_1}{1 + 5\lambda_2} & , \quad if \quad z_2 > 0 \, , \lambda_1 < |z_2| \\[2mm] \dfrac{z_2 + \lambda_1}{1 + 5\lambda_2} & , \quad if \quad z_2 < 0 \, , \lambda_1 < |z_2| \\[2mm] 0 & , \quad if \quad \qquad \lambda_1 \geq |z_2| \end{cases}$$

$$z_2 = \sum_{i=1}^{n} r_i \, x_{i2} + 4\lambda_2 \, \beta_3 - 2\lambda_2 \, \beta_1 - \beta_4$$

$$\text{For } j = p - 1, \hat{\beta}_{p-1} = \frac{S(z_{p-1}, \lambda_1)}{1 + 5\lambda_2} = \begin{cases} \dfrac{z_{p-1} - \lambda_1}{1 + 5\lambda_2} & , \quad if \quad z_{p-1} > 0 \, , \lambda_1 < |z_{p-1}| \\[2mm] \dfrac{z_{p-1} + \lambda_1}{1 + 5\lambda_2} & , \quad if \quad z_{p-1} < 0 \, , \lambda_1 < |z_{p-1}| \\[2mm] 0 & , \quad if \quad \qquad \lambda_1 \geq |z_{p-1}| \end{cases}$$

$$z_{p-1} = \sum_{i=1}^{n} r_i \, x_{ip-1} + 4\lambda_2 \, \beta_{p-2} - 2\lambda_2 \, \beta_p - \beta_{p-3}$$

$$\text{For } 3 \leq j \leq p, \quad \hat{\beta}_j = \frac{S(z_j, \lambda_1)}{1 + 6\lambda_2} = \begin{cases} \dfrac{z_j - \lambda_1}{1 + 6\lambda_2} & , \quad if \quad z_j > 0, \lambda_1 < |z_j| \\ \dfrac{z_j + \lambda_1}{1 + 6\lambda_2} & , \quad if \quad z_j < 0, \lambda_1 < |z_j| \\ 0 & , \quad if \quad \lambda_1 \geq |z_j| \end{cases}$$

$$z_j = \sum_{i=1}^{n} r_i x_{ij} + 4\lambda_2(\beta_{j+1} + \beta_{j-1}) - \lambda_2(\beta_{j+2} + \beta_{j-2}) \quad for \ 3 \leq j \leq p - 2$$

<u>Proof</u>:

The equation (4.7) is equivalent to the bellow

$$f(\beta^*) = \frac{1}{2} \sum_{i=1}^{n^*}(r_i^* - \beta_j^* x_{ij}^*)^2 + \lambda_1 \sum_{k \neq j}|\beta_k^*| + \lambda_1|\beta_j^*| \tag{4.8}$$

Where:

$$r_i^* = y_i^* - \sum_{k \neq j} x_{ik}^* \hat{\beta}_k^*, and \ n^* = n + p$$

The derivative of equation (4.8) with respect to β_j^* gives

$$\frac{df(\beta^*)}{d \ \beta_j^*} = \begin{cases} -\sum_{i=1}^{n^*} r_i^* x_{ij}^* + \beta_j^* \sum_{i=1}^{n^*} x_{ij}^{*2} + \lambda_1 & , \quad if \ \beta_j^* > 0 \\ -\sum_{i=1}^{n^*} r_i^* x_{ij}^* + \beta_j^* \sum_{i=1}^{n^*} x_{ij}^{*2} - \lambda_1 & , \quad if \ \beta_j^* < 0 \end{cases} \tag{4.9}$$

In the equation (4.9)

$$\sum_{i=1}^{n^*} r_i^* x_{ij}^* = \sum_{i=1}^{n} r_i x_{ij} + \sum_{i=1}^{p} r_i K_{ij} \tag{4.10}$$

$$\sum_{i=1}^{n^*} x_{ij}^{*2} = \sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{p} K_{ij}^2 \tag{4.11}$$

By plugging the equations (4.10) and (4.11) into (4.9), the derivatives become as

below:

$$\frac{df(\beta^*)}{d \ \beta_j^*} = \begin{cases} -\left(\sum_{i=1}^{n} r_i x_{ij} + \sum_{i=1}^{p} r_i K_{ij}\right) + \beta_j^*\left(\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{p} K_{ij}^2\right) + \lambda_1 & if \ \beta_j^* > 0 \\ -\left(\sum_{i=1}^{n} r_i x_{ij} + \sum_{i=1}^{p} r_i K_{ij}\right) + \beta_j^*\left(\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{p} K_{ij}^2\right) - \lambda_1 & if \ \beta_j^* < 0 \end{cases}$$

$$\tag{4.12}$$

Setting $\frac{df(\beta^*)}{d\ \beta_j^*} = 0$ and solving β_j^* gives five cases for β_j^* solutions

If ($j = 1$), then

$$For\ j = 1, \hat{\beta}_1^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{i1} + (-2\sqrt{\lambda_2}\,\beta_2^* + \sqrt{\lambda_2}\beta_3^*)(-\sqrt{\lambda_2}) - \lambda_1}{\sum_{i=1}^n x_{i1}^2 + (-\sqrt{\lambda_2})^2} & ,if\ \hat{\beta}_1^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{i1} + (-2\sqrt{\lambda_2}\,\beta_2^* + \sqrt{\lambda_2}\beta_3^*)(-\sqrt{\lambda_2}) + \lambda_1}{\sum_{i=1}^n x_{i1}^2 + (-\sqrt{\lambda_2})^2} & ,if\ \hat{\beta}_1^* < 0 \end{cases}$$

$$For\ j = p, \hat{\beta}_p^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ip} + (-2\sqrt{\lambda_2}\,\beta_{p-1}^* + \sqrt{\lambda_2}\beta_{p-2}^*)(-\sqrt{\lambda_2}) - \lambda_1}{\sum_{i=1}^n x_{ip}^2 + (-\sqrt{\lambda_2})^2} & ,if\ \hat{\beta}_p^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{ip} + (-2\sqrt{\lambda_2}\,\beta_{p-1}^* + \sqrt{\lambda_2}\beta_{p-2}^*)(-\sqrt{\lambda_2}) + \lambda_1}{\sum_{i=1}^n x_{ip}^2 + (-\sqrt{\lambda_2})^2} & ,if\ \hat{\beta}_p^* < 0 \end{cases}$$

For j=2

$$\hat{\beta}_2^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{i2} + (\sqrt{\lambda_2}\,\beta_1^* + \sqrt{\lambda_2}\beta_3^*)(2\sqrt{\lambda_2}) + (-2\sqrt{\lambda_2}\beta_3^* + \sqrt{\lambda_2}\beta_4^*) - \lambda_1}{\sum_{i=1}^n x_{i2}^2 + (2\sqrt{\lambda_2})^2 + (-\sqrt{\lambda_2})^2} & ,if\ \hat{\beta}_2^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{i2} + (\sqrt{\lambda_2}\,\beta_1^* + \sqrt{\lambda_2}\beta_3^*)(2\sqrt{\lambda_2}) + (-2\sqrt{\lambda_2}\beta_3^* + \sqrt{\lambda_2}\beta_4^*) + \lambda_1}{\sum_{i=1}^n x_{i2}^2 + (2\sqrt{\lambda_2})^2 + (-\sqrt{\lambda_2})^2} & ,if\ \hat{\beta}_2^* < 0 \end{cases}$$

For j=p-1

$$\hat{\beta}_{p-1}^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ip-1} + (\sqrt{\lambda_2}\,\beta_{p-3}^* - 2\sqrt{\lambda_2}\beta_{p-2}^*)(-\sqrt{\lambda_2}) + (\sqrt{\lambda_2}\beta_{p-2}^* + \sqrt{\lambda_2}\beta_p^*)(2\sqrt{\lambda_2}) - \lambda_1}{\sum_{i=1}^n x_{ip-1}^2 + (2\sqrt{\lambda_2})^2 + (-\sqrt{\lambda_2})^2} & ,\hat{\beta}_{p-1}^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{ip-1} + (\sqrt{\lambda_2}\,\beta_{p-3}^* - 2\sqrt{\lambda_2}\beta_{p-2}^*)(-\sqrt{\lambda_2}) + (\sqrt{\lambda_2}\beta_{p-2}^* + \sqrt{\lambda_2}\beta_p^*)(2\sqrt{\lambda_2}) + \lambda_1}{\sum_{i=1}^n x_{ip-1}^2 + (2\sqrt{\lambda_2})^2 + (-\sqrt{\lambda_2})^2} & ,\hat{\beta}_{p-1}^* < 0 \end{cases}$$

For $2 < j < p - 1$

$$\hat{\beta}_j^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ij} + 4\lambda_2 \left(\beta_{j+1}^* + \beta_{j-1}^*\right) - \sqrt{\lambda_2}\left(\beta_{j+2}^* + \beta_{j-2}^*\right) - \lambda_1}{\sum_{i=1}^n x_{ij}^2 + \left(-\sqrt{\lambda_2}\right)^2 + \left(2\sqrt{\lambda_2}\right)^2 + \left(-\sqrt{\lambda_2}\right)^2} & , if \ \hat{\beta}_j^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{ij} + 4\lambda_2 \left(\beta_{j+1}^* + \beta_{j-1}^*\right) - \sqrt{\lambda_2}\left(\beta_{j+2}^* + \beta_{j-2}^*\right) + \lambda_1}{\sum_{i=1}^n x_{ij}^2 + \left(-\sqrt{\lambda_2}\right)^2 + \left(2\sqrt{\lambda_2}\right)^2 + \left(-\sqrt{\lambda_2}\right)^2} & , if \ \hat{\beta}_j^* < 0 \end{cases}$$

As $\sum_{i=1}^n x_{ij}^2 = 1 \ \ for \ j = 1, 2, \dots, p$ the above cases simplify

For j= 1

$$\hat{\beta}_1^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{i1} + 2\lambda_2\,\beta_2^* - \lambda_2\beta_3^* - \lambda_1}{1 + \lambda_2} & , \quad if \ \ \hat{\beta}_1^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{i1} + 2\lambda_2\,\beta_2^* - \lambda_2\beta_3^* + \lambda_1}{1 + \lambda_2} & , \quad if \ \ \hat{\beta}_1^* < 0 \end{cases}$$

For j= p

$$\hat{\beta}_p^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ip} + 2\lambda_2\,\beta_{p-1}^* - \lambda_2\beta_{p-2}^* - \lambda_1}{1 + \lambda_2} & , \quad if \ \ \hat{\beta}_p^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{ip} + 2\lambda_2\,\beta_{p-1}^* - \lambda_2\beta_{p-2}^* + \lambda_1}{1 + \lambda_2} & , \quad if \ \ \hat{\beta}_p^* < 0 \end{cases}$$

For $j = 2$

$$\hat{\beta}_2^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{i2} + 4\,\lambda_2\,\beta_3^* + 2\lambda_2\,\beta_1^* - \lambda_2\,\beta_4^* - \lambda_1}{1 + 5\lambda_2} & , \quad if \ \ \ \hat{\beta}_2^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{i2} + 4\,\lambda_2\,\beta_3^* + 2\lambda_2\,\beta_1^* - \lambda_2\,\beta_4^* + \lambda_1}{1 + 5\lambda_2} & , \quad if \ \ \ \hat{\beta}_2^* < 0 \end{cases}$$

For $j = p - 1$

$$\hat{\beta}_{p-1}^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ip-1} + 4\,\lambda_2\,\beta_{p-2}^* + 2\lambda_2\beta_p^* - \lambda_2\beta_{p-3}^* - \lambda_1}{1 + 5\lambda_2} & , \quad if \ \ \hat{\beta}_{p-1}^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{ip-1} + 4\,\lambda_2\,\beta_{p-2}^* + 2\lambda_2\beta_p^* - \lambda_2\beta_{p-3}^* + \lambda_1}{1 + 5\lambda_2} & , \quad if \ \ \hat{\beta}_{p-1}^* < 0 \end{cases}$$

For $2 < j < p - 1$

$$\hat{\beta}_j^* = \begin{cases} \dfrac{\sum_{i=1}^n r_i x_{ij} + 4\lambda_2\left(\beta_{j+1}^* + \beta_{j-1}^*\right) - \sqrt{\lambda_2}\left(\beta_{j+2}^* + \beta_{j-2}^*\right) - \lambda_1}{1 + 6\lambda_2} & , \quad if \quad \hat{\beta}_j^* > 0 \\[4mm] \dfrac{\sum_{i=1}^n r_i x_{ij} + 4\lambda_2\left(\beta_{j+1}^* + \beta_{j-1}^*\right) - \sqrt{\lambda_2}\left(\beta_{j+2}^* + \beta_{j-2}^*\right) + \lambda_1}{1 + 6\lambda_2} & , \quad if \quad \hat{\beta}_j^* < 0 \end{cases}$$

By using the soft threshold operator, the solution is given by

$$For \ j = 1, \quad \hat{\beta}_1 = \frac{S(z_1, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_1 - \lambda_1}{1 + \lambda_2} & , \quad if \quad z_1 > 0 \, , \lambda_1 < |z_1| \\[3mm] \dfrac{z_1 + \lambda_1}{1 + \lambda_2} & , \quad if \quad z_1 < 0 \, , \lambda_1 < |z_1| \\[3mm] 0 & , \quad if \qquad\quad \lambda_1 \geq |z_1| \end{cases}$$

$$z_1 = \sum_{i=1}^n r_i x_{i1} + 2\lambda_2\, \beta_2 - \lambda_2 \beta_3$$

$$For \ j = p, \quad \hat{\beta}_p = \frac{S(z_1, \lambda_1)}{1 + \lambda_2} = \begin{cases} \dfrac{z_p - \lambda_1}{1 + \lambda_2} & , \quad if \quad z_p > 0 \, , \lambda_1 < |z_p| \\[3mm] \dfrac{z_p + \lambda_1}{1 + \lambda_2} & , \quad if \quad z_p < 0 \, , \lambda_1 < |z_p| \\[3mm] 0 & , \quad if \qquad\quad \lambda_1 \geq |z_p| \end{cases}$$

$$z_p = \sum_{i=1}^n r_i x_{ip} + 2\lambda_2\, \beta_{p-1} - \lambda_2 \beta_{p-2}$$

$$For \ j = 2, \quad \hat{\beta}_2 = \frac{S(z_2, \lambda_1)}{1 + 5\lambda_2} = \begin{cases} \dfrac{z_2 - \lambda_1}{1 + 5\lambda_2} & , \quad if \quad z_2 > 0 \, , \lambda_1 < |z_2| \\[3mm] \dfrac{z_2 + \lambda_1}{1 + 5\lambda_2} & , \quad if \quad z_2 < 0 \, , \lambda_1 < |z_2| \\[3mm] 0 & , \quad if \qquad\quad \lambda_1 \geq |z_2| \end{cases}$$

$$z_2 = \sum_{i=1}^n r_i\, x_{i2} + 4\lambda_2\, \beta_3 - 2\lambda_2\, \beta_1 - \beta_4$$

$$\text{For } j = p - 1, \hat{\beta}_{p-1} = \frac{S(z_{p-1}, \lambda_1)}{1 + 5\lambda_2} = \begin{cases} \dfrac{z_{p-1} - \lambda_1}{1 + 5\lambda_2} & , & if \quad z_{p-1} > 0, \lambda_1 < |z_{p-1}| \\[2mm] \dfrac{z_{p-1} + \lambda_1}{1 + 5\lambda_2} & , & if \quad z_{p-1} < 0, \lambda_1 < |z_{p-1}| \\[2mm] 0 & , & if \qquad \quad \lambda_1 \geq |z_{p-1}| \end{cases}$$

$$z_{p-1} = \sum_{i=1}^{n} r_i \, x_{ip-1} + 4\lambda_2 \, \beta_{p-2} - 2\lambda_2 \, \beta_p - \beta_{p-3}$$

$$\text{For } 3 \leq j \leq p, \quad \hat{\beta}_j = \frac{S(z_j, \lambda_1)}{1 + 6\lambda_2} = \begin{cases} \dfrac{z_j - \lambda_1}{1 + 6\lambda_2} & , & if \quad z_j > 0, \lambda_1 < |z_j| \\[2mm] \dfrac{z_j + \lambda_1}{1 + 6\lambda_2} & , & if \quad z_j < 0, \lambda_1 < |z_j| \\[2mm] 0 & , & if \qquad \quad \lambda_1 \geq |z_j| \end{cases}$$

$$z_j = \sum_{i=1}^{n} r_i \, x_{ij} + 4\lambda_2 \big(\beta_{j+1} + \beta_{j-1} \big) - \lambda_2 \big(\beta_{j+2} + \beta_{j-2} \big) \quad for \; 3 \leq j \leq p - 2$$

4.3. Computational Techniques

The $glmnet$ package in R program, written by (Jerome Friedman, Trevor Hastie and Rob Tibshirani 2008), contains very efficient procedures for fitting lasso or elastic-net regularization paths for generalized linear models. This algorithm is fast and can handle a large number of variables p. The efficiency of this algorithm comes from using cyclical coordinate descent in the optimization process [10]. Since the S-Lasso and SRF solve the lasso type problem with the augmented data set(X^*, y^*), shown in theorem 3 and theorem 4, the $glmnet$ algorithms can use for fitting the S-Lasso and SRF problems.

4.3.1 Tuning parameters

One of the most important functions to use in the $glmnet$ package is $cv.glmnet$. The $glmnet$ function is first run to get a sequence of λ-values that corresponding to

getting one additional non-zero coefficient. After getting the possibility λ_s, the program

does n-fold cross-validation with n=10 by default, so $glmnet$ is run n times, each with a

fraction $\frac{n-1}{n}$ of the data, and prediction error are collected on the remaining fold [29]. The

$elastic\ net$ penalty needs to compute a pair of parameters($\alpha\ and\ \lambda_1$), so for each α a

sequence of λ_1 is computed, then the best λ_1 is chosen that gives the minimum of the

mean cross-validation error (cvm). Also the S-Lasso and SRF have two parameters to

compute (λ_1 , λ_2), so the process first needs to use cross-validation to choose appropriate

values of the penalties $\lambda_1 and \lambda_2$. For each λ_2 a sequence of λ_1 is computed, then the

optimal λ_1 that gives the minimum value of the mean cross validation error [14].

CHAPTER FIVE: PERFORMANCE STUDY

In this section several simulated examples and a real data set will introduce to compute the mean square error (MSE) and mean prediction error (MPE) to present the results of Lasso, elastic net, S-Lasso and SRF problems.

5.1. Mean Square Error (MSE) and Mean Predictor Error (MPE)

There are two error terms to illustrate the experimental results of the lasso, elastic net, S-Lasso and SRF

1. Mean square error (MSE) is a measure of the quality of an estimator. In other words, MSE measures the expected squared distance between an estimator and the true underlying parameters

$$MSE = E(\hat{\beta} - \beta)^2 \tag{5.1}$$

2. The mean squared prediction error measures the expected squared distance between what a predictor predicts for a specific value and what the true value is, thus it is a measurement of the quality of predictor.

$$MPE = E\left\{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right\} \tag{5.2}$$

While $\beta's$ are supposed to be known in the simulated example, it is easy to compute the MSE, but in the real data set the MSE cannot be calculate because of the unknown $\beta's$, and instead of it, the MPE is calculated that closely related to MSE. To show relation between MSE and MPE suppose that

$$Y = \eta(x) + \varepsilon \tag{5.3}$$

Where:

$\eta(x) = X\beta$, $\varepsilon \sim N_n \, (\, 0, \sigma^2 \,)$

$(\hat{\eta}(x) = X\hat{\beta})$: is an estimate of $\eta(x)$

Therefore, mean square error is defined by:

$$MSE = E[\hat{\eta}(x) - \eta(x)]^2 \tag{5.4}$$

$$MPE = E[y - \hat{\eta}(x)]^2 = MSE + \sigma^2 \tag{5.5}$$

$$MPE = \; E[y - \hat{\eta}(x)]^2 = E[Y - \hat{\eta}(x) + \eta(x) - \eta(x)]^2$$

$$= \; E[\hat{\eta}(x) - \eta(x)]^2 + 2E\big[(\hat{\eta}(x) - \eta(x))(Y - \eta(x))\big] + E[Y - \eta(x)]^2$$

$E[\hat{\eta}(x) - \eta(x)]^2 = \sigma^2$, $E[Y - \eta(x)]^2 = MSE$, and $E\big[(\hat{\eta}(x) - \eta(x))(Y - \eta(x))\big] = 0$

Thus, $MPE = MSE + \sigma^2$

Therefore, the minimizing MPE is equivalent to minimizing MSE.

The MSE in term of matrix becomes [27].

$$MSE = (\hat{\beta} - \beta)^T \Sigma \, (\hat{\beta} - \beta) \tag{5.6}$$

Where

Σ: is the population covariance matrix of X.

5.2. Simulated data

In this section, six simulated examples are used to compare the prediction performance of the lasso, elastic net, S-Lasso and SRF. The first two examples were first used the in original paper of lasso to compare the prediction performance between the lasso and ridge regression [27]. Third and fourth examples were created as a grouped variable situation to show the prediction performance results between the lasso and

elastic net has been introduced in [34]. The last two examples are based on the

smoothness regression that have been stated in [15]

All examples are simulated from the true model

$$Y = \beta^T X + \sigma \, \varepsilon \qquad\qquad (5.7)$$

Where $\varepsilon \sim N(0,1)$

a)   In example 1, a data set is simulated with 20 observations and 8 predictors

$$\beta_j = 0.85 \qquad\qquad for \ j = 1, 2, \ldots, 8$$

The pairwise correlation between x_i and x_j is designed to be

$$\psi_{i,j} = 0.5^{|i-j|} \qquad\qquad for \ i, j \in \{1, 2, \ldots, 8\}, \text{ and } \sigma = 3$$

b)   In example 2, a data set is simulated with 30 observations and 40 predictors

$$\sigma = 3$$

$$\beta_j = \begin{cases} 3 & for \ j \in \{1, 2, \ldots, 15\} \\ 0 & otherwise \end{cases}$$

And the correlation variables are constructed

$$\psi_{i,i} = 1 \qquad\qquad for \ j \in \{1, 2, \ldots, 15\}$$

$$\psi_{i,j} = \frac{1}{(1 + 0.01)} \qquad\qquad for \ i \neq j, and \ (i, j) \in \{1, 2, \ldots, 15\}$$

$$\psi_{i,j} = 0 \qquad\qquad otherwise$$

c)   In example 3, the number of observations is 100, and the number of predictors is

40

$$\beta = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \qquad \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$$

$$\psi_{i,j} = 0.5 \qquad\qquad for \ i, j \in \{1, 2, \ldots, 40\}, \text{ and } \sigma = 15$$

d)   In example 4, the data set is simulated with 50 observations and 40 predictors

$$\beta = \underbrace{3,\ldots,3}_{15}, \underbrace{0,\ldots,0}_{25}) \text{ , and } \sigma = 15$$

The predictors X were generated as follows:

$$X_i = Z_1 + \varepsilon_i^x, \quad Z_1 \backsim N(0,1), \quad i = 1,\ldots,5,$$

$$X_i = Z_2 + \varepsilon_i^x, \quad Z_2 \backsim N(0,1), \quad i = 6,\ldots,10,$$

$$X_i = Z_3 + \varepsilon_i^x, \quad Z_3 \backsim N(0,1), \quad i = 11,\ldots,15,$$

$$X_i \backsim N(0,1) , \quad X_i \text{ independent identically distributed}, \quad i = 16,\ldots,40$$

Where ε_i^x are independent identically distributed $N(0,0.01), i = 1,\ldots,15$

e)   Example 5 is about Smooth regression vector. In that example, the regression

vector is be:

$$\beta_j = (3 - 0.2j)^2, \quad \text{for } j = 1,..,15$$

$$\beta_j = 0 \quad , \quad \text{otherwise}$$

The correlation between X_i and X_j is set by:

$$\psi_{i,j} = \exp(-|i - j|) \qquad \text{For } (i,j) \in \{1,..,p\}^2.$$

Two cases are tested with this example, the first one, when the $p < n$

$p = 50 , n = 70$. The second case is when $p > n$, $p = 100$ and $n = 30$.

f)   Example 6 is about high sparsity index and smooth regression vector where the

regression vector is designed by:

$$\beta_j = (4 - 0.1j)^2 \ , \qquad j \in \{1,..,40\}$$

$$\beta_j = 0 \qquad , \qquad \text{otherwise}$$

The correlations are the same as in example (e).

41

For this example difference number of p and n are tested in both cases when $p <$ n , and when $p > n.$

5.3. Results of Simulated Examples

In this section, the methods of the lasso, elastic net, S-Lasso, and SRF are compared with each other in terms of accuracy. The performance of their estimator $\hat{\beta}$ in term of mean square error (MSE) is clarified by box plots in Figures 1 to 9 and Tables 1 to 9.
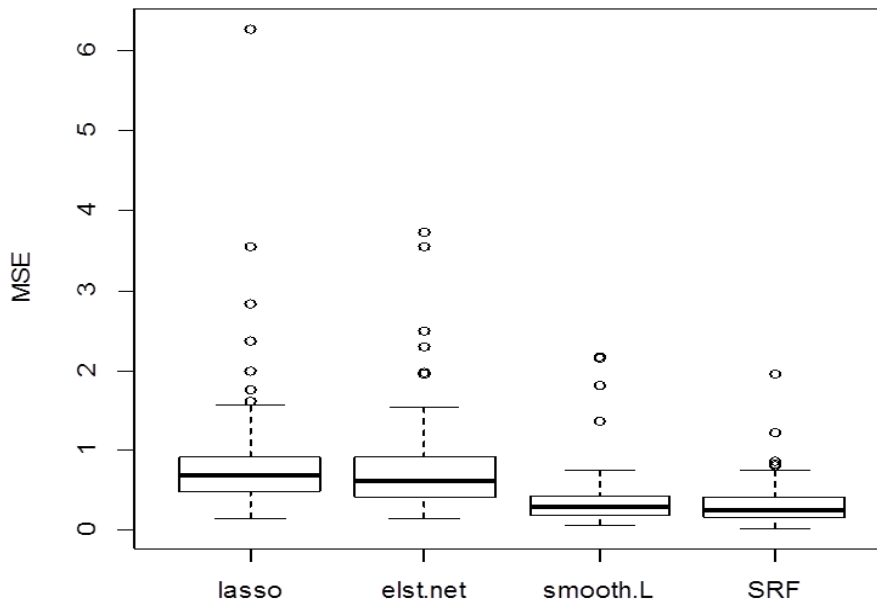

Figure 1: Comparing the accuracy of prediction of the lasso, the elastic net, the S-Lasso and the SRF, applied to Example (a) where n=20, p=8 and σ=3

Table 1: MSE for the simulated example (a) and number of nonzero coefficients of four methods where p=8 and n=20 (p<n).

Methods	Lasso	Elastic Net	S-Lasso	SRF
MSE	0.8521311	0.7885526	0.3766703	0.3264858
Non-zero Coefficients.	8	8	8	8

Table 1 and Figure 1 both summarize the predictor results of the simulated example (a).The results show that the mean squared error (MSE) value of the SRF (0.3264858) and the S-Lasso (0.3766703) are a little smaller than the lasso (0.8521311 and elastic net (0.7885526), the SRF and the S-Lasso therefore are significantly more accurate than the others. On the other hand, this example was constructed with some high correlated between covariates, therefore the lasso did not give the good results; the elastic net had a better performs than the lasso. In this example the number of predictors is less than the number of observations (p<n), and all of the four methods select all the variables without dropping any coefficient to 0.
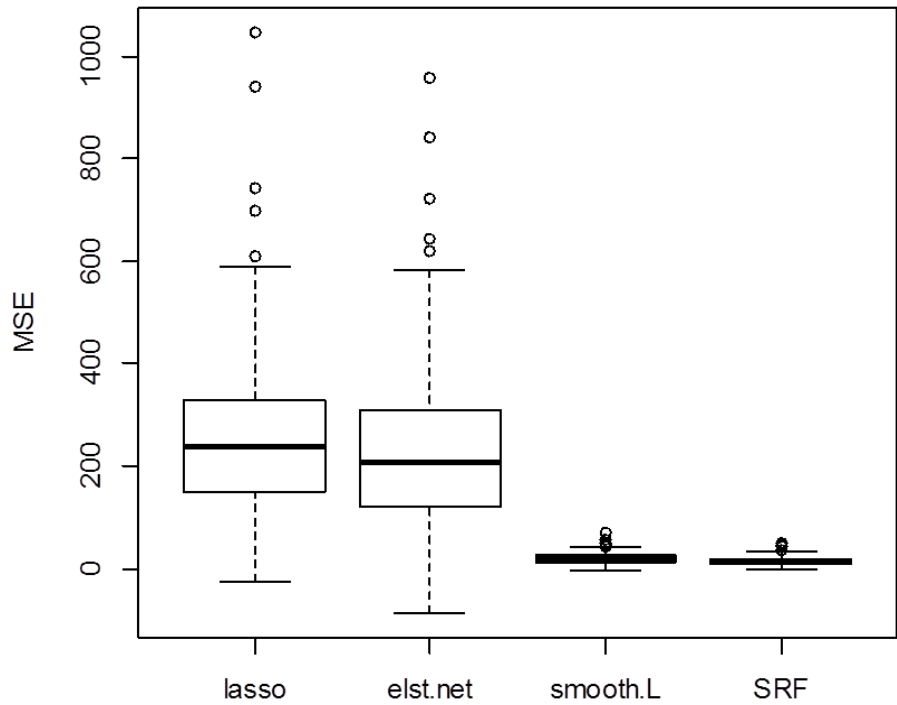
Figure 2:.Comparing the accuracy of prediction of the lasso, the elastic net, the S-Lasso and the SRF, applied to Example (b) where n=30, p=40 and σ=3.

Table 2: MSE for the simulated example (b) and number of nonzero coefficients of four methods where p=40 and n=30 (p<n).

Methods	Lasso	Elastic Net	S-Lasso	SRF
MSE	267.88064	236.77387	19.94284	15.59672
Non-zero	15	31	40	40
Coefficients.				

Table 2 and Figure 2 both summarize the predictor results of the simulated

example (b). This example is constructed with three groups of the correlated predictor

variables, and the results suggest that the SRF has a very small MSE value (15.59672)

compared to the lasso (267.88064) and elastic net (236.77387).Also, the S-Lasso gives a

good result with its MSE value (19.94284). This example was built with a situation of

some groups of highly correlated between covariate. It is clear that the lasso cannot give

good results under collinearity conditions, and the elastic net just gives small better

results than lasso in this situation, On the other hand, the S-Lasso designed to provide a

smooth and sparse solution, and this is true under the collinearity condition. However, the

S-Lasso cannot be better than the SRF because there are some points of roughness with

the true coefficients, and the SRF penalty is designed to penalize the roughness points

and makes successive coefficients close to each other. In this example the number of

predictors is more than the number of observations (p<n). Therefore, in selected

variables, lasso selected 15 variables out of the p=40 because lasso only selects one

variable from each group of highly correlated variables and ignored the rest while the

number of selected variables by the elastic net, the S-Lasso, and the SRF are 31, 40, and

40 respectively. Therefore, the SRF is significantly more accurate than the other methods.
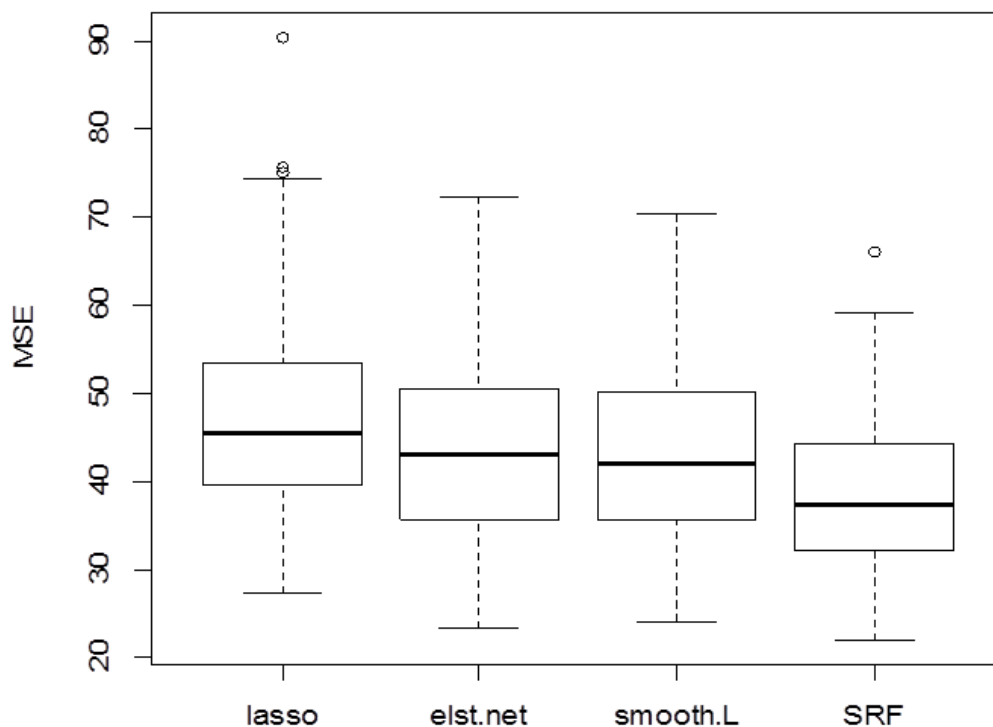
Figure 3:.Comparing the accuracy of prediction of the lasso, the elastic net, the S-Lasso and the SRF, applied to Example (c) where n=100, p=40 and σ=15.

The results of the example (c) are shown in the following table:

Table 3: MSE for the simulated example (c) and number of nonzero coefficients of four methods where p=40 and n=100 (p<n).

Methods	Lasso	Elastic Net	S-Lasso	SRF
MSE	47.29836	43.58239	43.02816	38.82478
Non-zero	21	22	21	22
Coefficients.				

Table 3 and Figure 3 show the predictor results of the simulated example (c). In this example the number of predictor variables is less than the number of observations ($p < n$). This example is constructed with an equal correlation between the covariates, and a set of 20 pure noise features; this means there are 20 true features and 20 noise features. The results show that the mean squared error (MSE) value of the SRF (38.82478) is smaller than the MSE values of the lasso (47.29836), elastic net (43.58239), and S-Lasso (43.02816) because the SRF solves the roughness of the coefficients. On the other hand, the S-Lasso gives better results than the lasso, and it presents almost the same results as elastic net because the coefficient vector is not smooth. Therefore, SRF is significantly more accurate than lasso, elastic net and S-Lasso. In the selected variables, the SRF and elastic net each selects 22 predictor variables and the others values of the coefficients are to 0 while the lasso and S-Lasso each selects 21 variables. Moreover, in this example true coefficients are set with two levels of values 0 and 3, and it gives some roughness points of coefficients. Therefore, the S-Lasso could not select more variables than the lasso in this situation. On the other hand, while there are not groups of highly correlated variables, it gives to lasso more ability to select variables. Results of the simulated example (c) demonstrate that the SRF outperforms of the lasso, elastic net, and the S-Lasso.
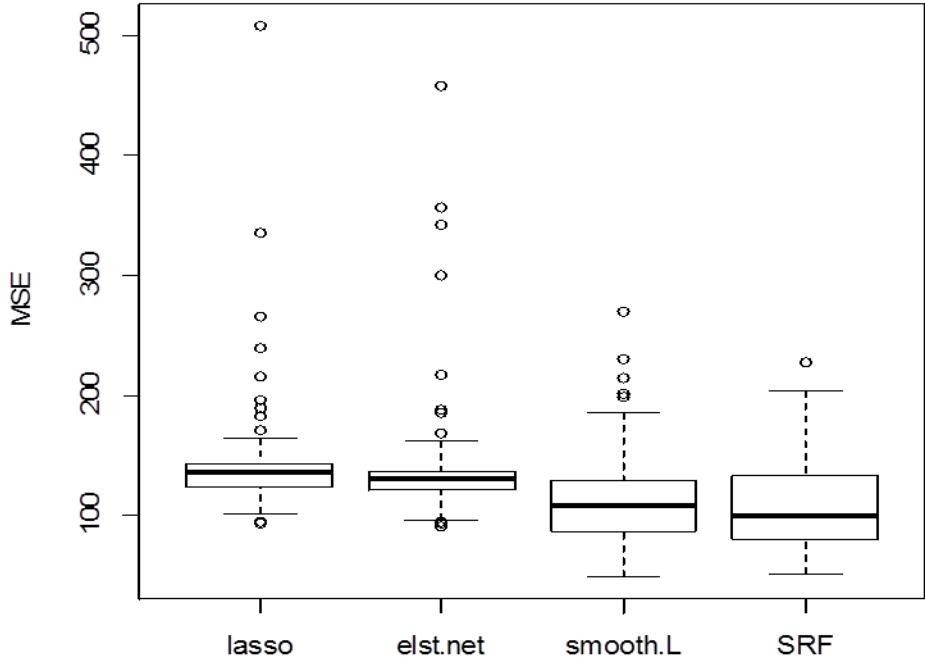
Figure 4:.Comparing the accuracy of prediction of the lasso, the elastic net, the S-Lasso and the SRF, applied to Example (d) where n=100, p=40 and σ=15

The results of the example (d) are shown in the following table:

Table 4: MSE for the simulated example (d) and number of nonzero coefficients of four   methods where p=40 and n=50 (p<n).

Methods	Lasso	Elastic Net	S-Lasso	SRF
MSE	141.7251	139.1600	111.7977	106.1705
Non-zero Coefficients.	11	12	17	20

Table 4 and Figure 4 show the results of the simulated example (d). This example contains three groups with five components for each groups, and a set of 25 pure noise features, it means there are 15 true features and 25 noise features. In this example, the number of predictor variables is less than the number of observations $(p < n)$. Since the mean squared error (MSE) value of the SRF (106.1705) is smaller than the MSE values of the other methods, the SRF is significantly more accurate than the lasso, elastic net and S-Lasso because the SRF tackles the roughness situation and the SRF always gives better results than the other methods under collinearity condition. On the other hand, the SRF selected 20 predictor variables out of $p = 40$ and gave 0 to the other coefficients while the number of variables selection by the lasso, elastic net and S-Lasso are (11, 12 and 17) respectively. While there are three groups of the covariates with strong collinearlity between the variables within each group, the lasso cannot be a good method to select variables. The elastic net gives better results with collinearity situation, but this example constructed with high dimension $(\sigma = 15)$, and the elastic net breaks down in selected variables under the high dimension situation.. Therefore, the results of the simulated example (d) show that the SRF dominates the methods of lasso, elastic net and S-Lasso.
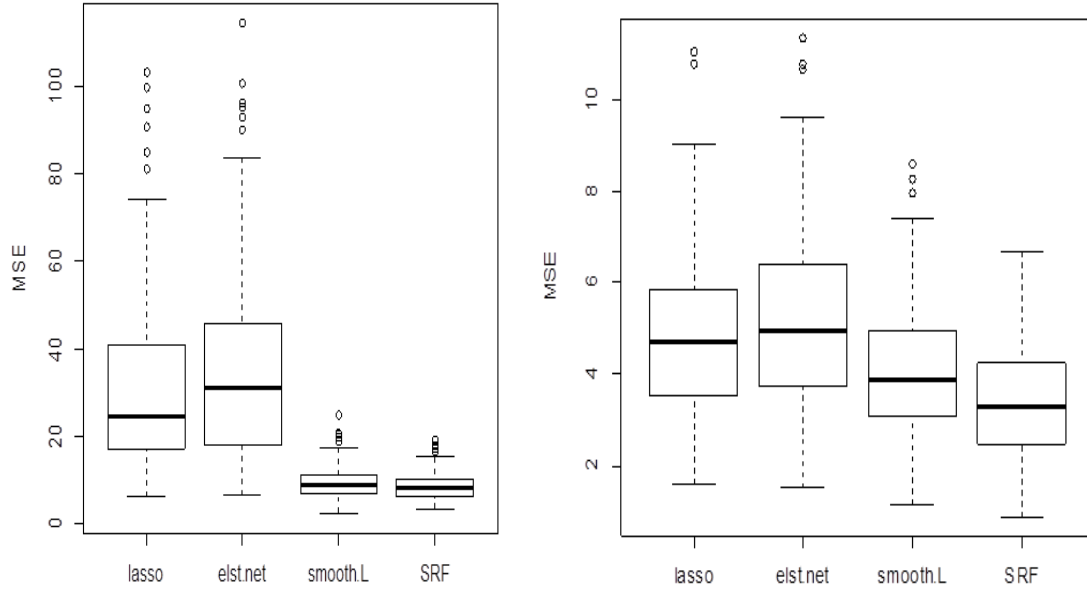
Figure 5:.Comparing the accuracy of prediction of the lasso, the elastic net, the S-Lasso and the SRF, applied to Example (e). Left plot is the case where p=100 and n =30 (p>n). Right plot is the case where p=50 and n=70 (p<n), σ=3 for both case.

The following tables are presented the results of example (e) with the both cases

p<n and p>n.

Table 5: MSE for the simulated example (e) and number of nonzero coefficients of four methods where p=50 and n=70 (p<n).

Methods	Lasso	Elastic Net	S-Lasso	SRF
MSE	4.898895	5.290685	4.1022717	3.495497
Non-zero	30	29	36	36
Coefficients.				

Table 6: MSE for the simulated example (d) and number of nonzero coefficients of four methods where p=100 and n=30 (p>n).

Methods	Lasso	Elastic Net	S-Lasso	SRF
MSE	31.009490	36.322612	9.464144	8.579100
Non-zero Coefficients.	20	30	66	64

Example (e) is the situation where the regression vector is smooth. In this example two cases are considered. The cases are related to the number of the observations n and predictors p. Right plot of the Figure 5 and Table 5 explain the results of the simulated example (e) with case of $p < n$ where $p = 50, n = 70$. There are two level set of true coefficients, a level with contains 15 components of nonzero and another one with 35 zero components. The coefficients of the regression vector vary slowly, or smooth. The results of the box plot and the table are clear that the mean squared error (MSE) value of the SRF (3.495497) is a smaller than the MSE values of the other methods, therefore SRF is significantly more accurate than the lasso, elastic net and S Lasso. It is clear that the S-Lasso also gives a better performance than the lasso and elastic net under the smooth regression vector, but the S-Lasso doesn't gives better results than the SRF because the SRF penalizes the roughness within three points while the S-Lasso does it in two points. On the other hand, while there is a big group of zero components, the elastic net has a poor performance compared with the lasso. In the case of variable selection, the SRF and S-Lasso both select 36 predictor variables out of $p = 50$ and gave 0 to the other coefficients. The number of variables selected by the

lasso and elastic net are 30 and 29 respectively. Therefore, the results of the simulated example (e) demonstrate that the SRF dominates the methods of the lasso, the elastic net and the S-Lasso.

Left plot of Figure 5 and Table 6 show the results of the simulated example (e) in case of p>n where p=100 and n=30. According to the results, the SRF has a high performance compare with the other methods, and the lasso still performs better than the elastic net. Here the elastic net still has poor results because of big set of the zero features. On the other hand, the SRF and S-Lasso give a big set of non-zero coefficients 64 and 66 respectively which means both have more predictor variables than sample size n. The lasso gives only 20 non-zero coefficients and elastic net gives 30, and it is clear that the lasso selected variables less than the sample size. Therefore, the results of the simulated example (e) in both cases show that the SRF outperforms the methods of lasso, elastic net and S-Lasso.
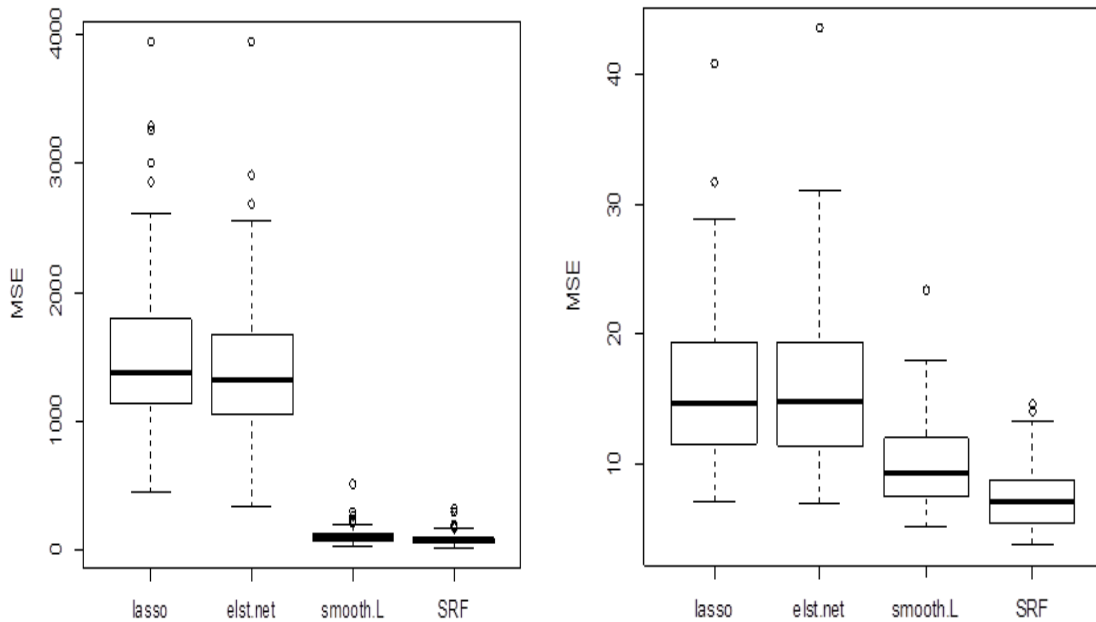
Figure 6: Comparing the accuracy of prediction of the lasso, the elastic net, the S-Lasso and the SRF, applied to Example (f). Left plot is the case where p=100 and n =30 (p>n). Right plot is the case where p=50 and n=70 (p<n), σ=3 for both case.

The results of the both cases (p<n) and (p>n) of the example (e) are shown in the

two following tables

Table 7: MSE for the simulated example (f) and number of nonzero coefficients of four methods where p=50 and n=70 (p<n).

Methods	Lasso	Elastic Net	S-Lasso	SRF
MSE	15.77443	15.92286	10.02195	7.46686
Non-zero Coefficients.	43	45	45	45

Table 8: MSE for the simulated example (f) and number of nonzero coefficients of four methods where p=100 and n=30 (p>n).

Methods	Lasso	Elastic Net	S-Lasso	SRF
MSE	1533.27022	1394.21921	113.88964	86.22573
Non-zero Coefficients.	12	17	55	60

Example (f) is the situation of high sparsity index and smooth regression vector. In this example two cases are discussed. When the number of predictors p is less than the number of observations n $(p < n)$ and the opposite situation when $(p > n)$. Left plot of the Figure 6 and Table 7 summarize the results of the simulated example (f) with case of $p < n$ where $p = 50, n = 70$. The results show that the mean squared error (MSE) value of the SRF (7.46686) is smaller than the MSE of the other methods, and SRF is significantly more accurate than the lasso, elastic net and S-Lasso. The S-Lasso also has a better performance compared than the lasso and elastic net because of the smoothness of the regression vector. On the other hand, while the MSE value of the lasso (15.77443) is smaller than the MSE value of the elastic net (15.92286), the lasso outperforms the elastic net because in this case the sparsity index=40 is smaller than the sample size n=70. The SRF, the S-Lasso and the elastic net each selects 45 predictor variables out of $p = 50$ and gave 0 for the other coefficients while the number of variables selected by the lasso is 43.

Right plot of Figure 6 and Table 8 explain the results of the simulated example (f) in case of p>n where p=100 and n=30. According to the results, the SRF has a high performance compare with other methods. Also the S-Lasso has better performance than

the lasso and elastic net. In the smoothness of the regression vector, the SRF gives the

better performance the S-Lasso. On the other hand, in this case (p>n) the elastic net

performed better than the lasso because the sparsity index is more than the sample size in

this case. The SRF and S-Lasso give a big set of the number of non-zero coefficients 60

and 55 respectively which means both had more predictor variables than sample size n

while the lasso gives only 12 non-zero coefficients and elastic net gives 17. Therefore,

the results of the simulated example (f) in both cases show that the SRF dominates the

methods of lasso, elastic net and S-Lasso.

5.4. Real Data Set

In this section the calibration problems in near-infrared (NIR) is applied which

considered in [2]. The original spectral data consist of 700 points measured from 1100 to

2498 nanometers (nm) wavelengths in steps of 2 nm. Four constituents of biscuit dough

fat, sugar, flour, and water discussed in this prediction. They had 40 calibration set or

training set and 40 validation set measured at a different time, but they excluded one

sample from each of these two data sets, leaving 39 samples in each. The problem was to

use these spectra to predict four constituents of the dough-fat, sugar, flour, and water. In

this thesis, the data sets focus only on the constituent of the dough- water.

Since $\beta's$ are unknown with the real data set, the MSE cannot be easily calculated.

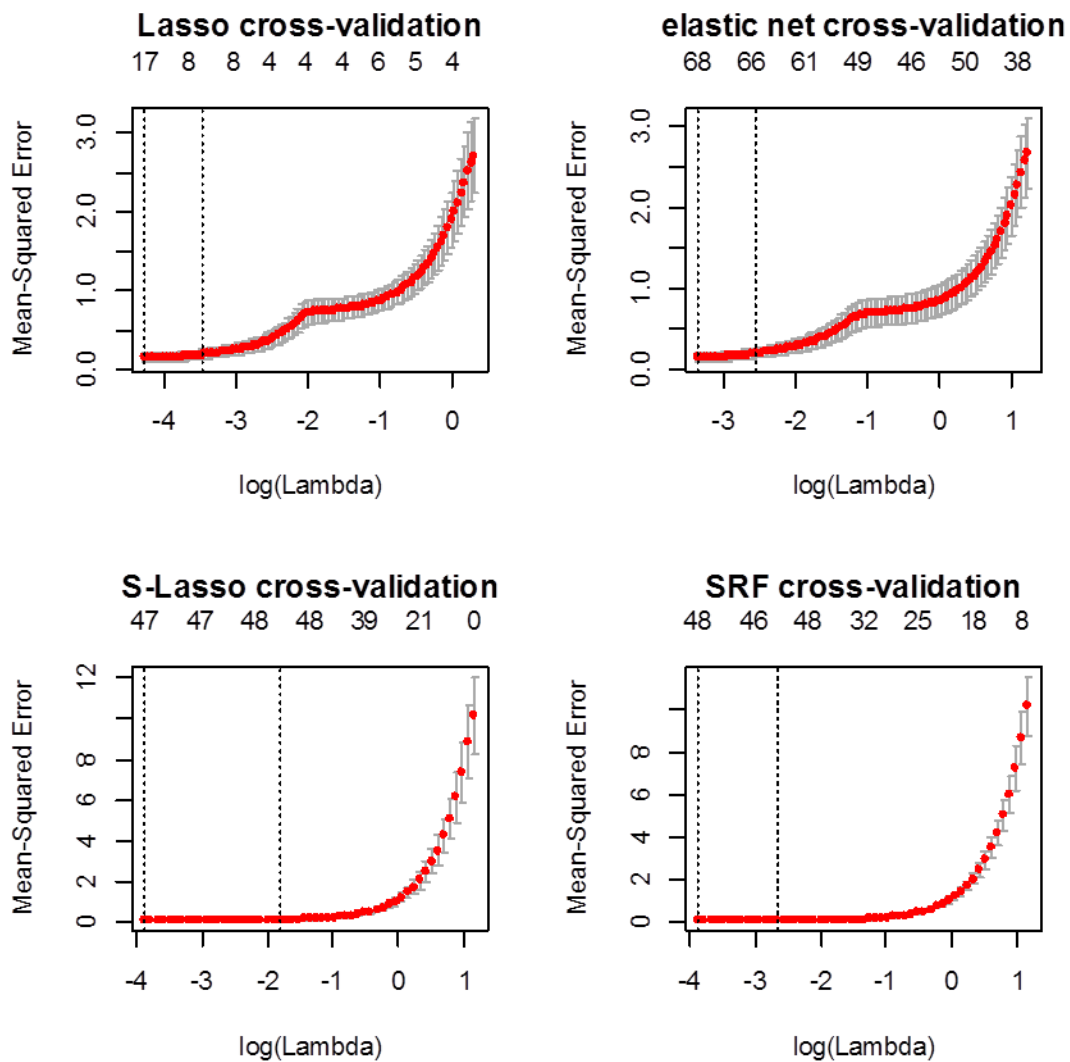However, predicted error (MPE) can be calculated, which is closely related to MSE.

Figure 7:.Evaluation of cross-validation plots of the lasso, the elastic net, the S-Lasso and the SRF, based on the calibration data (dough-water) to choose the   best lambda that gives the minimum MSE.

It is clear from the C.V plot in Figure 7 that as lambda increases, the MSE increases rapidly. The coefficients are reduced too much and they do not adequately fit the responses. In contrast, as lambda decreases, the models are larger and have more nonzero coefficients. The increasing MSE suggests that the models are over-fitted.
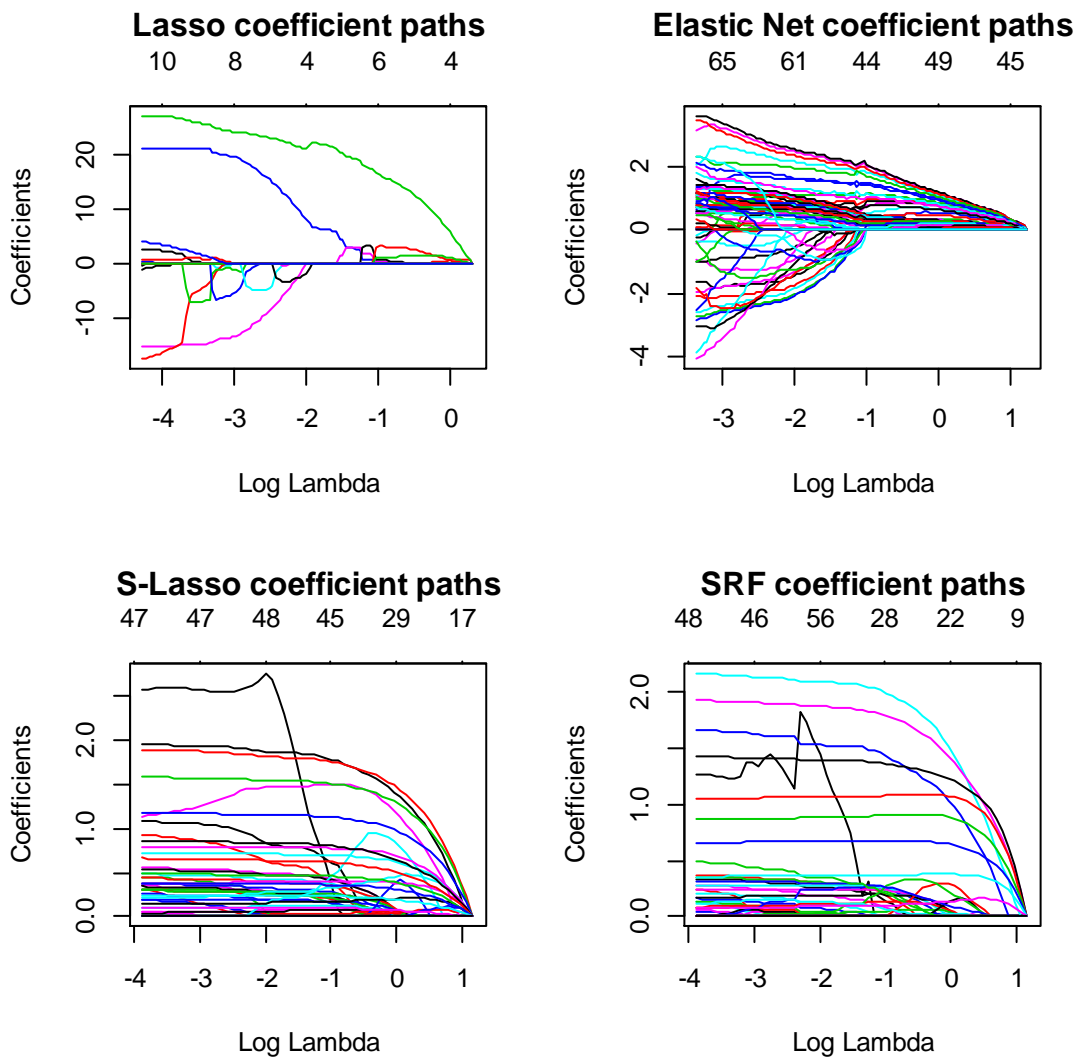
Figure 8: plots of the number of predictors in the fitted lasso, elastic net, S-lasso and SRF regularization as a function of lambda.

The graph of the lasso, elastic net, S-Lasso and SRF on the Figure 8 estimate which variables enter the model based on the lambda of their estimates. The optimal solution depends on the selected value of lambda, which is chosen based on cross-validation.
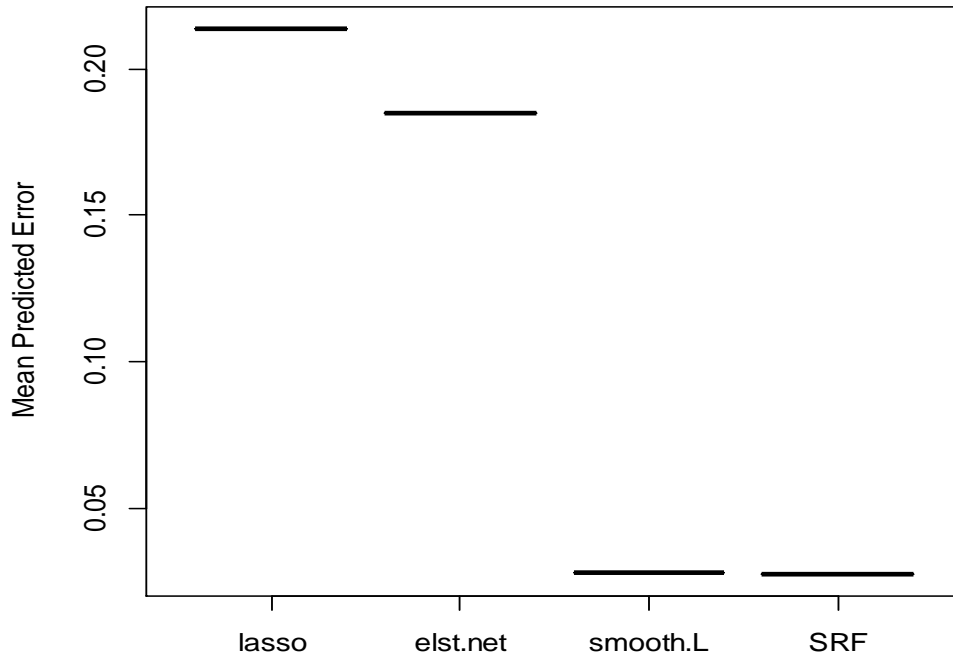
Figure 9:.Comparing the accuracy of prediction of the lasso, the elastic net, the S-Lasso and the SRF, applied to calibration data (dough-water) where p=700 and the number observations based on the training and validation sets, n=39 for each of the sets

The following table is summarized the results:

Table 9: MPE for the calibration data set (dough-water) and number of nonzero coefficients of four methods where p=700 and n=39 (p>n).

Methods	Lasso	Elastic Net	S-Lasso	SRF
MPE	0.21358440	0.18511157	0.02819653	0.02778441
Non-zero Coefficients.	9	66	49	47

The plot of Figure 9 and Table 9 explain the results of the calibration data set focused only on dough-water. The situation here is the number of predictors $(p = 700)$ greatly exceeds the number of observations $(n = 39)$. The procedure of the fitted model depends on two stages. The first stage which is based on the training data set and the model is fitted in this stage. Then for stage two the validation data set is used to test the results. According to the results, the mean predictor error (MPE) value of the SRF (0.02778441) is smaller than the other methods results; therefore SRF is significantly more accurate than the lasso, elastic net and S-Lasso. On the other hand, the SRF, the S-Lasso, and the elastic net selected more predictor variables than sample size n, which are 47, 49 and 66 respectively while the lasso only selected 8 variables and gave 0 coefficients to others. Therefore, the SRF overcomes the methods of the lasso, elastic net and S-Lasso.

CHAPTER SIX: COUNCLUSIONS

This thesis proposed sparse ridge fusion (SRF) as a new procedure to solve the generalized linear model problem. Several simulated examples with different situations of the grouping effect and smooth regression vector were used to test the SRF and compared with the lasso, elastic net and smooth lasso methods. The conclusions demonstrate that the SRF appears to perform well on all simulated examples compare to the other methods in terms of prediction accuracy in the cases when $p \leq n$ and when $p > n$. Moreover, when the methods are tested on the calibration data set (dough-water), the results illustrate that the SRF dominates the lasso, elastic net, and S-Lasso.

LIST OF REFRENCES

1. Berkson, J. (1944). Application of the logistic function to bio-assay. Journal of the American Statistical Association, 39(227), 357-365.

2. Brown, P., Fearn, T., & Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. Journal of the American Statistical Association, 96(454), 398-408.

3. Bunea, F. (2008a). Consistent selection via the Lasso for high dimensional approximating regression models. Institute of Mathematical Statistics Collections, 3, 122-137.

4. Bunea, F. (2008b). Honest variable selection in linear and logistic regression models via $\ell 1$ and $\ell 1 + \ell 2$ penalization. Electronic Journal of Statistics, 2, 1153-1194.

5. Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association, 90(432), 1200-1224.

6. Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. The Annals of Statistics, 32(2), 407-499.

7. Elmer, S. G. (2011). Modern statistical methods applied to economic time series. KOF Swiss Economic Institute, ETH Zurich, Zürich. Available from http://worldcat.org /z-wcorg/ database.

8. Flexeder, C. (2010). Generalized Lasso Regularization for Regression Models. Institut für Statistik.

9. Fox, J. (2000a). Multiple and generalized nonparametric regression: Sage.

10. Friedman, J., Hastie, T., & Tibshirani, R. (2008). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), 1.

11. Galton, F. (1886). Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

12. Hastie, T., & Loader, C. (1993). Local regression: Automatic kernel carpentry. Statistical Science, 8(2), 120-129.

13. Hastie, T., Tibshirani, R., & Friedman, J. (2009). Linear Methods for Regression: Springer.

14. Hawkins, D. M., & Maboudou-Tchao, E. M. (2013). Smoothed Linear Modeling for Smooth Spectral Data. International Journal of Spectroscopy, 2013.

15. Hebiri, M., & van de Geer, S. (2011). The Smooth-Lasso and other $\ell 1 + \ell 2$-penalized methods. Electronic Journal of Statistics, 5, 1184-1226.

16. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

17. Jia, J., & Yu, B. (2008). On model selection consistency of the Elastic Net when P>> n: DTIC Document.

18. Kutner, M. H., Nachtsheim, C., & Neter, J. (2004). Applied linear regression models.

19. Land, S., & Friedman, J. (1996). Variable fusion: a new method of adaptive signal regression: Technical Report.

20. Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. Electronic Journal of Statistics, 2, 90-102.

21. Mays, J. E., Birch, J. B., & Alden Starnes, B. (2001). Model robust regression: combining parametric, nonparametric, and semiparametric methods. Journal of Nonparametric Statistics, 13(2), 245-277.

22. Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, 34(3), 1436-1462.

23. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821): Wiley.

24. Neter, J., Wasserman, W., & Kutner, M. H. (1996). Applied linear statistical models (Vol. 4): Irwin Chicago.

25. Ogutu, J., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings, 6(Suppl 2), S10.

26. Ozanne, M., & Dyar, M. D. (2012). Comparison of Shrunken Regression Methods for Major Elemental Analysis of Rocks Using Laser-Induced Breakdown Spectroscopy (LIBS). Mount Holyoke College.

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

28. Wainwright, M. J. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. arXiv preprint math/0605740.

29. Weisberg, S. 1980. Applied Linear Regression. Wiley, New York.

30. Wright, D. B., & London, K. (2009). Modern regression techniques using R: a practical guide: Sage.

31. Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. The Annals of Applied Statistics, 224-244.

32. Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. The Journal of Machine Learning Research, 7, 2541-2563.

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

34. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320. doi: 10.1111/j.1467-9868.2005.00503.x

35. Zou, H., & Zhang, H. H. (2009). On the adaptive elastic- net with a diverging number of parameters. Annals of Statistics, 37(4), 1733.