A Comparative Study of Ridge, LASSO and Elastic net Estimators

by

Meaad Aldabal

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of
Master of Science

in

Statistics

Carleton University
Ottawa, Ontario

Abstract

The focus of this thesis is to review the three basic penalty estimators, namely, ridge regression estimator, LASSO, and elastic net estimator in the light of the deficiencies of least-squares estimator. Ill-conditioned design matrix is the major source of problem in this case. To overcome this problem, ridge regression was developed, and it opened the door for penalty estimators. Its impact is visible with various linear and non-linear models. A superb discovery in the class of subset selection is the LASSO (Least Absolute Shrinkage and Selection Operator) which selects subsets and estimates the coefficients simultaneously. Finally, we consider the elastic net penalty estimator which combine the L_1 and L_2 penalty function. Resulting estimator is weighted LASSO by ridge factor. We obtain the L_2-risk expressions and compare with pre-test and Stein-type estimators. For the location model, we discovered that the naive elastic net is better than elastic net estimators as opposed to the conclusion in the current literature. On the other hand in case of regression model, the elastic net performs reasonably compared to LASSO and ridge regression.

Keywords: risk function, efficiency, lasso, penalty estimator, ridge estimator, pre-test and Stein-type estimators, elastic net estimator, tuning parameter.

Acknowledgements

I would like to thank my parents and my family for encouraging me to step out of my comfort zone to go to a foreign nation to pursue higher studies. Without their love and affection, I would not have been able to complete this research project. Next, I would like to thank the Saudi Arabia Ministry of Education for providing scholarships that enabled me to fund my education and living expenses here in Canada, which motivated me to keep pursuing my joy and passion for learning. Most importantly, Professor Saleh deserves my heartfelt gratitude and thanks for being patient with me, inspiring me to do research at the highest level, and for allowing me to learn at my required pace to have a good understanding of the concepts. His commitment towards learning and education continues to amaze me, and who despite his age strives forward elegantly to prove new results and invent new methods. Many thanks to Carleton University for providing me with this wonderful opportunity and learning experience, part of which includes the wonderful faculty, staff, and students.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

The history of estimation method has kept on advancing through the years beginning with Gauss (1795) who proposed the basic "Least-squares" methodology published by Adrine-Marie Legendre (1805). Next, R.A Fisher (1922) introduced the method of maximum likelihood and included variety of properties such as consistency, sufficiency, efficiency and information.

Stein (1961) put forward a fundamental result which states: Stein estimator of mean vector of a multivariate normal distribution of dimension three or more, dominates the maximum likelihood estimator which was expanded by Efron (1973) by means of the empirical Bayes method. Later, Saleh and Sen (1978 - 1987) expanded Stein's idea to embrace rank-based theory of statistics contributing to the area of "robust methodology". Another notable methodology is the "ridge regression" put forward by Hoerl and Kennard (1970) that opened the door for "penalty estimation". Impact of penalty estimators is now visible in data analysis with regression models. A superb discovery in this class of estimators is the LASSO (Least Absolute Shrinkage and Selection Operator) by Tibshirani (1996). It is an effective method of subset selection of the coefficients of regression models and estimation thereof, which is the most appropriate efficient way of model building via data analysis. Data analysis usually makes use of least-squares (LS) estimate and predicts response based on the LS estimators. However, data analysts are not satisfied with the results they get out of the LS method, even though the estimators are unbiased with minimum variance. To understand the shortcomings of using LS estimation, we consider the model with the data $\{(\boldsymbol{x}_i, y_i)|i = 1, \ldots, n\}$ where $\boldsymbol{x}_i$ is the ith row of the X matrix given in (1.1) below:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i \ ; \ \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)', \ \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \ 1 < i < n. \tag{1.1}$$

The least-square estimator of $\boldsymbol{\beta}$ is $\tilde{\boldsymbol{\beta}}_n = (X'X)^{-1} X'Y$, where $\mathrm{E}(\tilde{\boldsymbol{\beta}}_n) = \boldsymbol{\beta}$ and $\mathrm{Cov}(\tilde{\boldsymbol{\beta}}_n) = \sigma^2 (X'X)^{-1}$. Let X^0 is a future X-matrix, then the predicted value of Y would be $\hat{Y} = X^0 \tilde{\boldsymbol{\beta}}_n$. We have $\mathrm{Cov}(\hat{Y}) = \sigma^2 X^{0\prime} (X'X)^{-1} X^0$. Since σ^2 is unknown, it can be estimated by

$$s^2 = (n-p)^{-1} \left(Y - X\tilde{\boldsymbol{\beta}}_n\right)' \left(Y - X\tilde{\boldsymbol{\beta}}_n\right).$$

Now, we consider the result carefully. First, look at the prediction accuracy: LS estimate has no bias but may have high variance for some estimator of coefficients. This would spoil the accuracy of prediction. The second reason is the interpretation: Data analysts prefer a simpler model because it sheds more light on the relationship between the response and covariates. Thus, subset selection is important when the number of predictors is large, so that their number can reasonably be reduced for optimal interpretation.

To improve the LS methodology, one uses LASSO subset selection introduced by Tibshirani (1996), ridge regression introduced by Hoerl and Kennard (1970) and elastic net, which is a combination of LASSO and ridge regression penalty suggested by Zou and Hastie (2005). We shall focus on these three subset selectors in this thesis. We may consider the traditional subset selector, namely, the preliminary test subset selector (PTSS) proposed by Donoho and Johnstone (1994) as:

$$\hat{\boldsymbol{\beta}}_n^{PT}(\lambda) = \left(\tilde{\beta}_{1n} \ I\left(|\tilde{\beta}_{1n}| > \lambda\sigma\sqrt{c^{11}}\right), \ldots, \tilde{\beta}_{pn} \ I\left(|\tilde{\beta}_{pn}| > \lambda\sigma\sqrt{c^{pp}}\right)\right)', \tag{1.2}$$

where c^{jj} is j^{th} diagonal element of $(X'X)^{-1} = C_n^{-1}$, and $I(A)$ is the indicator function of the set A, i.e. $I(A)$ takes a value of 1 if A occurs and takes a value of 0 if A^c occurs.

This subset selector in (1.2) can be highly variable because it is a discrete process. Here, regression coefficients are either retained or dropped from the model. Small change in the data can result in very different models being selected and this can effect their prediction accuracy.

As a result, one may consider the continuous version of PTSS, and propose Saleh-type subset selector (SSS) as

$$\hat{\boldsymbol{\beta}}_n^S(\lambda) = \left(\tilde{\beta}_{1n}\left(1 - \frac{\lambda\sigma\sqrt{c^{11}}}{|\tilde{\beta}_{1n}|}\right), \ldots, \tilde{\beta}_{pn}\left(1 - \frac{\lambda\sigma\sqrt{c^{pp}}}{|\tilde{\beta}_{pn}|}\right)\right)'. \tag{1.3}$$

The problem with (1.3) is the fact that it changes the sign of the estimator.

To modify the selector in (1.3), we consider the positive-rule of Saleh-type selector, namely,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^{S+}(\lambda) = &\left(\tilde{\beta}_{1n}\left(1 - \frac{\lambda\sigma\sqrt{c^{11}}}{|\tilde{\beta}_{1n}|}\right) I\left(|\tilde{\beta}_{1n}| > \lambda\sigma\sqrt{c^{11}}\right), \ldots,\right.\\ &\left.\tilde{\beta}_{pn}\left(1 - \frac{\lambda\sigma\sqrt{c^{pp}}}{|\tilde{\beta}_{pn}|}\right) I\left(|\tilde{\beta}_{pn}| > \lambda\sigma\sqrt{c^{pp}}\right)\right)'.\end{aligned} \tag{1.4}$$

The left hand-side of (1.4) form may be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^{S+}(\lambda) = &\left(sgn\left(\tilde{\beta}_{1n}\right)\left(|\tilde{\beta}_{1n}| - \lambda\sigma\sqrt{c^{11}}\right)^+, \ldots,\right.\\ &\left.sgn\left(\tilde{\beta}_{pn}\right)\left(|\tilde{\beta}_{pn}| - \lambda\sigma\sqrt{c^{pp}}\right)^+\right)'\\ &= \sigma\left(\sqrt{c^{11}}\; sgn\left(Z_{1n}\right)\left(|Z_{1n}| - \lambda\right)^+, \ldots, \sqrt{c^{pp}}\; sgn\left(Z_{pn}\right)\left(|Z_{pn}| - \lambda\right)^+\right)',\end{aligned} \tag{1.5}$$

where $Z_{jn} = \frac{\tilde{\beta}_{jn}}{\sigma\sqrt{c^{jj}}}, j = 1, \ldots, p$ and (1.5) is the LASSO selector expression. One may notice that the last expression does not change sign.

Now, we present the naive elastic net (nEnet) selector given by

$$\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda, \alpha) = \left(\frac{\hat{\beta}_{1n}^{LASSO}(\lambda\alpha)}{1 + \lambda\left(1 - \alpha\right)\; c^{11}}, \ldots, \frac{\hat{\beta}_{pn}^{LASSO}(\lambda\alpha)}{1 + \lambda\left(1 - \alpha\right)\; c^{pp}}\right)', \tag{1.6}$$

and $0 \leq \alpha \leq 1$ is the elastic tuning parameter.

From the formula (1.6), when we set $\alpha = 0$, we get the usual ridge estimator

$$\tilde{\boldsymbol{\beta}}_n^{ridge}(\lambda) = \left(\frac{\tilde{\beta}_{1n}}{1+\lambda c^{11}}, \ldots, \frac{\tilde{\beta}_{pn}}{1+\lambda c^{pp}} \right)',$$

and for $\alpha = 1$, we obtain the LASSO

$$\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda) = \left(\hat{\beta}_{1n}^{LASSO}(\lambda), \ldots, \hat{\beta}_{pn}^{LASSO}(\lambda) \right)'.$$

The thesis has five chapters according to the materials covered. Chapter 2 contains introduction to ridge regression, mainly focussed on linear regression models. location model is considered only to illustrate the properties under L_2-risk criterion. In particular, we discuss application of partial penalty function when model indicates sparsity for the model and includes application of Stein-type estimators. We present the basic analytical methodology. So that one may use them in any problem related to ridge regression. It is well-known, ridge regression performs efficiently enough in data analysis. It is a shrinkage estimator but does not select any variables. LASSO is brought in to simultaneously select and estimate a coefficient to build efficient models for prediction. This is presented in chapter 3. In chapter 4, we consider elastic net penalty function which is a combination of L_1- and L_2-penalty functions. This results in a weighted LASSO by ridge factor i.e.

$$\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda, \alpha) = [I_p + kC^{-1}]^{-1} \hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha) \ , \ 0 \leq \alpha \leq 1 \ , \ k = \lambda(1-\alpha).$$

These results are presented in chapter 4 and show that Enet is worse than nEnet. Chapter 5 contains the concluding remarks.

Chapter 2

RIDGE REGRESSION

In this chapter, we present some basic results of the ridge regression estimation with their important properties. Ridge regression introduced by Hoerl and Kennard (1970), is a breakthrough contribution and a competitive estimation procedure to the Stein-type estimators. We derive the ridge regression estimator and study its risk properties. To begin with, we consider two models, namely, the location model and the multiple linear regression (MLR) model in the following sub sections.

2.1 Some Basic Information

Let us consider the location model

$$Y_n = \theta 1_n + \epsilon \; ; \quad \mathrm{E}(\epsilon) = \underline{0}, \; \mathrm{E}(\epsilon\epsilon') = \sigma^2 I_n, \; \sigma^2 \text{ known}, \tag{2.1}$$

where $Y_n = (y_1, \ldots, y_n)'$, 1_n-is the n-vector of 1's and θ is unknown parameter. In order to have the best linear unbiased estimator (BLUE) for (2.1) form, we have

Theorem 2.1.1.

$$\tilde{\theta}_n = \arg\min_{\theta \in R} (Y_n - \theta 1_n)' (Y_n - \theta 1_n) = \bar{y}. \tag{2.2}$$

Proof. Taking derivative of $(Y_n - \theta 1_n)' (Y_n - \theta 1_n)$ with respect to θ and setting to 0, we get $\tilde{\theta}_n = \bar{y}$ as shown in (2.2).

□

Clearly, $\mathrm{E}\left(\tilde{\theta}_n\right) = \theta$ and $\mathrm{Var}(\tilde{\theta}_n) = \dfrac{\sigma^2}{n}$.

To get an improved estimator of θ, we use the penalized least-squares loss-function.

Theorem 2.1.2.

$$\tilde{\theta}_n^{ridge} = \arg\min_{\theta}\{(Y_n - \theta 1_n)'(Y_n - \theta 1_n) + n\lambda\theta^2\} = \frac{\tilde{\theta}}{1+\lambda}. \tag{2.3}$$

Proof. Taking derivative with respect to θ and setting it to 0, we solve for θ to get

$$\tilde{\theta}_n^{ridge} = \frac{\tilde{\theta}}{1+\lambda},$$

which represents the formula (2.3).

□

It can be easily shown that

$$\mathrm{E}\left(\tilde{\theta}_n^{ridge}\right) = \frac{\theta}{1+\lambda}, \ \mathrm{Bias}\left(\tilde{\theta}_n^{ridge}\right) = \frac{\theta}{1+\lambda} - \theta = -\frac{\lambda\theta}{1+\lambda}.$$

Then, the risk of $\tilde{\theta}_n^{ridge}(\lambda)$ is given by (2.4) as follows

$$\begin{aligned} R\left(\tilde{\theta}_n^{ridge}(\lambda);\theta\right) &= \mathrm{Var}\left(\tilde{\theta}_n^{ridge}(\lambda)\right) + \mathrm{Bias}^2\left(\tilde{\theta}_n^{ridge}(\lambda)\right) \\ &= \frac{\sigma^2}{n}\frac{1}{(1+\lambda)^2} + \frac{\lambda^2\theta^2}{(1+\lambda)^2} \\ &= \frac{\sigma^2}{n(1+\lambda)^2}\left(1+\lambda^2\Delta^2\right), \ \Delta^2 = \frac{n\theta^2}{\sigma^2}. \end{aligned} \tag{2.4}$$

The optimum value for λ is Δ^{-2}. Hence, $R\left(\hat{\theta}_n^{ridge}(\Delta^{-2});\theta\right) = \frac{\sigma^2\Delta^2}{[n(1+\Delta^2)]} < \frac{\sigma^2}{n}$ for all $\Delta^2 > 0$. Therefore, under squared loss $\hat{\theta}_n^{ridge}(\Delta^{-2})$ is uniformly better than the best linear unbiased estimator (BLUE).

2.2 Linear Model

Consider the following multiple linear regression model,

$$Y = \theta 1_n + X\boldsymbol{\beta} + \epsilon, \tag{2.5}$$

where

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}_{n\times p} ; \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} ; Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1},$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, and θ is the intercept and random error $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$. Let $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)'$ where $\boldsymbol{x}_i$ is the ith row of X. For the time being, assume $Y = X\boldsymbol{\beta} + \epsilon$ with the following standard assumptions:
1) $\mathrm{E}(\epsilon) = 0$,
2) $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ for all values of X, where σ^2 is known,
3) $\epsilon = (\epsilon_1, \dots, \epsilon_n)' \approx N_n(0, \sigma^2 I_n)$, I_n is the identity matrix.

Theorem 2.2.1. *The best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is*

$$\tilde{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in R^p} \{(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta})\} = (X'X)^{-1} X'Y, \tag{2.6}$$

provided $X'X$ is not ill conditioned. Then, we have $\mathrm{Var}\left(\tilde{\boldsymbol{\beta}}_n\right) = \sigma^2 (X'X)^{-1}$.

Proof. To prove the formula (2.6), we set the derivative of $(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta})$ to 0 to get:

$$-2X'(Y - X\boldsymbol{\beta})' = 0 \Rightarrow \tilde{\boldsymbol{\beta}}_n = (X'X)^{-1} X'Y.$$

□

2.2.1 Reparametrization of the Design Matrix X

In this section, we consider the canonical form of the multiple linear regression model (2.5) to obtain the expressions for L_2-risks of estimators in

terms of eigenvalues of $X'X$ matrix. Let $C = X'X$ be a positive definite matrix. So, there exists an orthogonal matrix Γ such that $X'X = \Gamma\Lambda\Gamma'$ and $\Lambda = \Gamma'C\Gamma = \text{diag}\,(\lambda_1, \ldots, \lambda_p)$, where $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$ are the ordered eigen values of $X'X$.

Then, the new parametrization of the multiple linear regression model (MLR) is given by (2.7) as follows

$$Y = T\boldsymbol{\xi} + \epsilon, \quad T = X\Gamma, \quad \boldsymbol{\xi} = \Gamma'\boldsymbol{\beta}, \tag{2.7}$$

and the least squares estimator $\tilde{\boldsymbol{\xi}}_n$ and its Cov-matrix together with risk expressions are given by

$$\tilde{\boldsymbol{\xi}}_n = \Lambda^{-1}T'Y,$$
$$\text{Cov}\left(\tilde{\boldsymbol{\xi}}_n\right) = \sigma^2\Lambda^{-1}, \quad R\left(\tilde{\boldsymbol{\xi}}_n; \boldsymbol{\xi}\right) = \sigma^2 \,\text{tr}\,\Lambda^{-1} = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}.$$

2.3 Ridge Regression Estimator

Theorem 2.3.1. *Under the above assumptions the ridge regression estimator is given by*

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_n^{ridge}(k) &= \arg\min_{\boldsymbol{\beta}\in R^p} \left[(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}) + k\boldsymbol{\beta}'\boldsymbol{\beta}\right] \\ &= \left[X'X + k\ I_p\right]^{-1} X'Y.\end{aligned}$$

Proof. Setting the derivative of $\left[(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta})\right] + k\boldsymbol{\beta}'\boldsymbol{\beta}$ with respect to k equal to 0, we obtain

$$\left[X'X + kI\right]\boldsymbol{\beta} = X'Y.$$

Hence, the ridge regression estimator of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}}_n^{ridge}(k) = \left[C_n + kI_p\right]^{-1} X'Y,$$

where $C_n = X'X$.

□

Now, we have the following lemma.

Lemma 2.1:

$$\begin{aligned}
\mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right) &= \sigma^2 \left[C_n + kI\right]^{-1} C_n \left[C_n + kI\right]^{-1}, \\
\mathrm{Bias}\left(\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right) &= -k \left[C_n + kI\right]^{-1} \boldsymbol{\beta}, \\
R\left(\tilde{\boldsymbol{\beta}}_n^{ridge}(k) : \boldsymbol{\beta}\right) &= \sigma^2 \,\mathrm{tr}\{\left[C_n + kI\right]^{-2} C_n\} + k^2 \boldsymbol{\beta}' \left[C_n + kI\right]^{-2} \boldsymbol{\beta}.
\end{aligned}$$

Proof.

$$\begin{aligned}
\mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right) &= \left[C_n + kI\right]^{-1} X' \mathrm{E}\left[YY'\right] X \left[C_n + kI\right]^{-1} \\
&= \sigma^2 \left[C_n + kI\right]^{-1} C_n \left[C_n + kI\right]^{-1}. \\
\mathrm{Bias}\left(\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right) &= \left[C_n + kI\right]^{-1} \left[X'X\boldsymbol{\beta} - k\boldsymbol{\beta} - X'X\boldsymbol{\beta}\right] \\
&= -k \left[C_n + kI\right]^{-1} \boldsymbol{\beta}.
\end{aligned}$$

To prove the expression for risk of the ridge regression in canonical form, we consider the eigenvalues of $C_n + kI$, which are $\lambda_1 + k, \ldots, \lambda_p + k$ and the eigenvalues of $\left(C_n + kI\right)^{-1} C_n$ are $\frac{\lambda_1}{(\lambda_1+k)^2}, \ldots, \frac{\lambda_p}{(\lambda_p+k)^2}$, respectively by (2.7). Then, we can derive the risk as follows:

$$\begin{aligned}
R\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(k); \boldsymbol{\xi}\right) &= \mathrm{E}\left[\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(k) - \boldsymbol{\xi}\right)' \left(\tilde{\boldsymbol{\xi}}_n^{ridge}(k) - \boldsymbol{\xi}\right)\right] \\
&= \sum_{j=1}^{p} \mathrm{E}\left(\tilde{\xi}_j^{ridge}(k) - \xi_j\right)^2 \\
&= \sum_{j=1}^{p} \left[\mathrm{Var}\left(\tilde{\xi}_j^{ridge}(k)\right) + k^2 \,\mathrm{Bias}^2\left(\tilde{\xi}_j^{ridge}(k)\right)\right] \\
&= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \sum_{j=1}^{p} \frac{\xi_j^2}{(\lambda_j + k)^2} \\
&= \sigma^2 \,\mathrm{tr}\{(\Lambda + kI)^{-2}\Lambda\} + k^2 \boldsymbol{\xi}' \left(\Lambda + kI\right)^{-2} \boldsymbol{\xi}.
\end{aligned}$$

Note that

$$\mathrm{Var}\left(\tilde{\xi}_j^{ridge}(k)\right) = \sigma^2 \frac{\lambda_j}{(\lambda_j + k)^2}, \;\; \mathrm{Bias}\left(\tilde{\xi}_j^{ridge}(k)\right) = -k \frac{\xi_j}{(\lambda_j + k)}.$$

□

Now, we have the following theorem

Theorem 2.3.2. *There always exists a* $k > 0$ *in the range* $0 < k < \frac{\sigma^2}{\xi^2_{max}}$ *such that* $\tilde{\boldsymbol{\xi}}_n^{ridge}(k)$ *has a smaller risk than* $\tilde{\boldsymbol{\xi}}_n$. *i.e.*

$$R\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(k); \boldsymbol{\xi}\right) \leq R\left(\tilde{\boldsymbol{\xi}}_n; \boldsymbol{\xi}\right) \text{ uniformly in } \boldsymbol{\xi} \in R^p.$$

Proof. To prove the theorem, we consider the risk expression for $\tilde{\boldsymbol{\xi}}_n^{ridge}(k)$ as a function of k. If $k = 0$, we get $\sigma^2 \sum \frac{1}{\lambda_j}$. To illustrate, the first term appears as a continuous and monotonic decreasing function of k. The derivative of this function (with respect to k) tends to -∞ when $k \to 0^+$. The second term in the expression is a continuous and monotonic increasing function. Second term approaches $\boldsymbol{\xi}'\boldsymbol{\xi}$ when $k \to \infty$. Differentiating with respect to k we get

$$\frac{dR}{dk}\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(k) : \boldsymbol{\xi}\right) = 2\sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^3}\left(k\xi_j^2 - \sigma^2\right). \tag{2.8}$$

Then, a sufficient condition for (2.8) to be negative is that $0 < k < k^*$, where

$$k^* = \frac{\sigma^2}{\xi^2_{max}} = \Delta^{-2}_{max}; \quad \xi^2_{max} = \max\left(\xi_1^2, \ldots, \xi_p^2\right).$$

□

2.3.1 Generalized Ridge Regression Estimator of $\boldsymbol{\beta}$

In this section, we consider the generalized ridge regression estimator of $\boldsymbol{\beta}$ using penalized least-square function i.e

$$\tilde{\boldsymbol{\beta}}_n(K) = \arg\min_{\boldsymbol{\beta} \in R^p} \left\{(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}) + \boldsymbol{\beta}' K \boldsymbol{\beta}\right\}, \; K = \text{diag}\left(k_1, \ldots, k_p\right). \tag{2.9}$$

Setting the derivative equal to 0, we obtain (2.9) given by

$$\left[X'X + K\right] \boldsymbol{\beta} = X'Y \;\Rightarrow\; \tilde{\boldsymbol{\beta}}_n^{ridge}(K) = \left[X'X + K\right]^{-1} X'Y. \tag{2.10}$$

If $K = kI_p$, then the equation (2.10) will be written as follows

$$\tilde{\boldsymbol{\beta}}_n^{ridge}(k) = [X'X + kI_p]^{-1} X'Y.$$

The following theorem shows that the ridge regression estimator performs better than the LSE $\left(\tilde{\boldsymbol{\beta}}_n\right)$.

Theorem 2.3.3. *Under the assumed conditions*

$$\begin{aligned} R\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(K); \boldsymbol{\xi}\right) &= \sigma^2 \operatorname{tr}\left\{[\Lambda + K]^{-2} \Lambda\right\} + \boldsymbol{\xi}' K [\Lambda + K]^{-2} K\boldsymbol{\xi}, \\ R\left(\tilde{\boldsymbol{\xi}}_n; \boldsymbol{\xi}\right) &= \sigma^2 \operatorname{tr} \Lambda^{-1}, \\ R\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(K); \boldsymbol{\xi}\right) &\leq R\left(\tilde{\boldsymbol{\xi}}_n; \boldsymbol{\xi}\right) \ \forall\ \boldsymbol{\xi} \in R^p. \end{aligned}$$

Proof. We consider the canonical form and note that,

$$\begin{aligned} \operatorname{Cov}\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(K)\right) &= \sigma^2 (\Lambda + K)^{-1} \Lambda (\Lambda + K)^{-1}, \\ \operatorname{Bias}\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(K)\right) &= -[\Lambda + K]^{-1} K\boldsymbol{\xi}, \\ \operatorname{Bias}^2\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(K)\right) &= \boldsymbol{\xi}' K [\Lambda + K]^{-2} \ K\boldsymbol{\xi}, \\ R\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(K); \boldsymbol{\xi}\right) &= \sigma^2 \ \left[\sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k_j)^2} \ + \sum_{j=1}^{p} \frac{k_j^2 \Delta_j^2}{(\lambda_j + k_j)^2}\right], \end{aligned}$$

where $\Delta_j^2 = \frac{\xi_j^2}{\sigma^2},\ j = 1, \ldots, p.$

□

It may be shown that the optimum value for k_j is $\Delta_j^{-2};\ (j = 1, \ldots, p)$. Hence, the optimum value for

$$R\left(\tilde{\boldsymbol{\xi}}_n^{ridge}(K) : \boldsymbol{\xi}\right) = \sum_{j=1}^{p} \frac{\sigma^2}{\left(\lambda_j + \frac{1}{\Delta_j^2}\right)} < \sum_{j=1}^{p} \frac{\sigma^2}{\lambda_j} \ \forall \left(\Delta_1^2, \ldots, \Delta_p^2\right) \in R^p.$$

This proves that the ridge regression estimator dominates $\tilde{\boldsymbol{\xi}}_n$ uniformly in $\boldsymbol{\xi} \in R^p$ under L_2-loss function.

Now, we consider the case when $p > n$. In this case $X'X$ is a $p \times p$ matrix with minimum rank n, which means we cannot obtain the inverse of $X'X$.

We resolve the problem by considering artificial dataset (X^*, Y^*) defined by

$$X^* = \begin{pmatrix} X_{n\times p} \\ \sqrt{k} I_p \end{pmatrix}_{(n+p)\times p} \quad , \quad Y^* = \begin{pmatrix} Y \\ 0 \end{pmatrix}_{(n+p)\times 1}.$$

Thus, the multiple linear model is given by

$$Y^*_{(n+p)\times 1} = X^*_{(n+p)\times p}\, \boldsymbol{\beta}_{p\times 1} + \epsilon_{(n+p)\times 1}.$$

Then, LS estimator is given by

$$\tilde{\boldsymbol{\beta}}^* = (X^{*\prime} X^*)^{-1}\ X^{*\prime} Y^*,$$

where

$$\begin{pmatrix} X' & \sqrt{k} I_p \end{pmatrix} \begin{pmatrix} X \\ \sqrt{k} I_p \end{pmatrix} = \left[X'X + kI_p\right],$$

$$X^{*\prime}\, Y = X'Y,$$

which result in the same LS estimator of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}}^{ridge}(k) = \left[X'X + kI_p\right]^{-1} X'Y,$$
$$\tilde{\boldsymbol{\xi}}^{ridge}(k) = \left(\Lambda + kI_p\right)^{-1} T'Y.$$

2.3.2 Estimation of $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)'$ when $\boldsymbol{\beta}_2$ may be sparse

In this section, we consider the estimation of the parameter $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ when we suspect $\boldsymbol{\beta}_2$ may be negligible for the model

$$Y = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \epsilon.$$

In this case, we minimize

$$(Y - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2)'\,(Y - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2) + k\ \boldsymbol{\beta}_2'\boldsymbol{\beta}_2\ ,$$

to get the following equations

$$\left\{\begin{array}{rcl} X_1'X_1\boldsymbol{\beta}_1 + X_1'X_2\boldsymbol{\beta}_2 & = & X_1'Y, \\ X_2'X_1\boldsymbol{\beta}_1 + (X_2'X_2 + kI_p)\ \boldsymbol{\beta}_2 & = & X_2'Y. \end{array}\right.$$

We obtain the following

$$\left(\begin{array}{c} \tilde{\boldsymbol{\beta}}_{1n} \\ \tilde{\boldsymbol{\beta}}_{2n} \end{array}\right) = \left(\begin{array}{cc} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 + kI_p \end{array}\right)^{-1} \left(\begin{array}{c} X_1'Y \\ X_2'Y \end{array}\right)$$

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k) &= \left[X_1'M_2X_1 + kI_{p_1}\right]^{-1} X_1'M_2Y, \\ \tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k) &= \left[X_2'M_1X_2 + kI_{p_2}\right]^{-1} X_2'M_1Y, \\ M_1 &= I_n - X_1\left(X_1'X_1\right)^{-1} X_1', \\ M_2 &= I_n - X_2\left(X_2'X_2 + kI_{p_2}\right)^{-1} X_2'\ , \end{aligned} \tag{2.11}$$

respectively. The above estimates are important in the analysis of high dimensional problem see Gao et al. (2017) and Saleh et al. (2019).
The bias and risk expressions are given by

$$\begin{aligned} \text{Bias}\left(\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k)\right) &= \left[X_1'M_2X_1 + kI_{p_1}\right]^{-1} X_1'M_2\left[X_1'\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2\right] - \boldsymbol{\beta}_1 \\ &= -k\left[X_1'M_2X_1 + kI_{p_1}\right]^{-1}\left[\boldsymbol{\beta}_1 + k^{-1}\left(X_1'M_2X_2\right)\boldsymbol{\beta}_2\right], \end{aligned}$$

$$\begin{aligned} \text{Bias}\left(\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k)\right) &= \left[X_2'M_1X_2 + kI_{p_2}\right]^{-1}\left[\left(X_2'M_1X_1\right)\boldsymbol{\beta}_1 - \left(X_2'M_1X_2\right)\boldsymbol{\beta}_2\right] - \boldsymbol{\beta}_2 \\ &= -k\left[X_2'M_1X_2 + kI_{p_2}\right]^{-1}\boldsymbol{\beta}_2. \end{aligned}$$

$$\begin{aligned} R\left(\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k);\boldsymbol{\beta}_1\right) &= \sigma^2\,\text{tr}\left\{\left[X_1'M_2X_1 + kI_{p_1}\right]^{-2}\left(X_1'M_2^2X_1\right)\right\} \\ &\quad + k^2\left[\boldsymbol{\beta}_1 - k^{-1}\left(X_1'M_2X_2\right)\boldsymbol{\beta}_2\right]'\left[X_1'M_2X_1 + kI_{p_1}\right]^{-2} \\ &\quad \left[\boldsymbol{\beta}_1 - k^{-1}\left(X_1'M_2X_2\right)\boldsymbol{\beta}_2\right]\ , \end{aligned}$$

$$\begin{aligned} R\left(\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k);\boldsymbol{\beta}_2\right) &= \sigma^2\,\text{tr}\left\{\left[X_2'M_1X_2 + kI_{p_2}\right]^{-2}\left(X_2'M_1X_2\right)\right\} \\ &\quad + k^2\boldsymbol{\beta}_2'\left[X_2'M_1X_2 + kI_{p_2}\right]^{-2}\boldsymbol{\beta}_2. \end{aligned}$$

Further, for the test of $H_0 : \boldsymbol{\beta}_2 = 0$ Vs $\boldsymbol{\beta}_2 \neq 0$, we use the test-statistics

$$\mathcal{L}_n = \frac{1}{\hat{\sigma}_n^2}\ \tilde{\boldsymbol{\beta}}_{2n}'\left(X_2'M_1X_2\right)\tilde{\boldsymbol{\beta}}_{2n}\ ,$$

where $M_1^2 = M_1$ and $\hat{\sigma}_n^2 = (n-p_2)^{-1}\left(Y - X_2\tilde{\boldsymbol{\beta}}_{2n}\right)'\left(Y - X_2\tilde{\boldsymbol{\beta}}_{2n}\right)$.

We can now, define the following estimators:
(1) Unrestricted estimator (UE):

$$\tilde{\boldsymbol{\beta}}_n^{UE}(k) = \left(\left[\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k)\right]',\ \left[\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k)\right]'\right)'.$$

(2) Preliminary test Estimator (PTE):

$$\hat{\boldsymbol{\beta}}_n^{PT}(k) = \left(\left[\tilde{\boldsymbol{\beta}}_{1n}^{ridge}\right]',\ \tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k)\ I\left(\mathcal{L}_n > k^2\right)\right)'.$$

(3) Stein-Saleh type Estimator (S):

$$\tilde{\boldsymbol{\beta}}_n^{S}(k) = \left(\left[\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k)\right]',\ \left[\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k)\right]'\left(1-(p_2-2)\mathcal{L}_n^{-1}\right)\right)'.$$

(4) Positive-rule Stein-Saleh type estimator (PRSS):

$$\tilde{\boldsymbol{\beta}}_n^{S+}(k) = \left(\left[\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k)\right]', \left[\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k)\right]'\left(1-(p_2-2)\mathcal{L}_n^{-1}\right) I\left(\mathcal{L}_n > p_2-2\right)\right)'.$$

We can prove the following results using the ideas of Saleh (2006).

(1) $R\left(\tilde{\boldsymbol{\beta}}_n^{S+}(k);\boldsymbol{\beta}\right) \leq R\left(\tilde{\boldsymbol{\beta}}_n^{S}(k);\boldsymbol{\beta}\right) \leq R\left(\tilde{\boldsymbol{\beta}}_n^{UE}(k);\boldsymbol{\beta}\right)$ uniformly in $\boldsymbol{\beta} \in R^p$.
(2) Neither $\tilde{\boldsymbol{\beta}}_n^{PT}(k)$ nor $\tilde{\boldsymbol{\beta}}_n^{S+}(k)$, $\tilde{\boldsymbol{\beta}}_n^{S}(k)$ or $\tilde{\boldsymbol{\beta}}_n^{UE}(k)$ dominate the other uniformly.

Now, consider the case with penalty $\boldsymbol{\beta}_2'\boldsymbol{\beta}_2$. Then,

$$\tilde{\boldsymbol{\beta}}_n^{ridge}(k) = \left(X'X + kI_p\right)^{-1}X'Y,$$

$$\begin{aligned}\begin{pmatrix} \tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k) \\ \tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k) \end{pmatrix} &= \left[\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} + k\begin{pmatrix} I_{p_1} & 0 \\ 0 & I_{p_2} \end{pmatrix}\right]^{-1}\begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix} \\ &= \left[\begin{pmatrix} X_1'X_1 + kI_{p_1} & X_1'X_2 \\ X_2'X_1 & X_2'X_2 + kI_{p_2} \end{pmatrix}\right]^{-1}\begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix}.\end{aligned}$$

Then,

$$\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k) = \left[X_1'M_2X_1 + kI_{p_1}\right]^{-1} \; X_1'M_2Y,$$
$$\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k) = \left[X_2'M_1^*X_2 + kI_{p_2}\right]^{-1} \; X_2'M_1^*Y;$$
$$M_1^* = I_n - X_1\left(X_1'X_1 + kI_{p_2}\right)^{-1} \; X_1',$$
and M_2 is the same as (2.11).

Hence, we may obtain the risk of these estimators as

$$\begin{aligned} R\left(\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k); \boldsymbol{\beta}_1\right) &= \sigma^2 tr\left\{\left[X_1'M_2X_1 + kI_{p_1}\right]^{-2}\left(X_1'M_2^*X_1\right)\right\} \\ &\quad + k^2\boldsymbol{\beta}_1'\left[X_1'M_2X_1 + kI_{p_1}\right]^{-2}\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2'\left(X_1'M_2X_2\right) \\ &\quad \left[X_1'M_2X_2 + kI_{p_1}\right]^{-2} \; \left(X_1'M_2X_2\right) \; \boldsymbol{\beta}_2. \end{aligned}$$

Similarly,

$$\begin{aligned} R\left(\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k); \boldsymbol{\beta}_2\right) &= \sigma^2 tr\left\{\left[X_2'M_1^*X_2 + kI_{p_2}\right]^{-2}\left(X_2'M_1^{*2}X_2\right)\right\} \\ &\quad + k^2\boldsymbol{\beta}_2'\left[X_2'M_1^*X_1 + kI_{p_2}\right]^{-2}\boldsymbol{\beta}_2 + \boldsymbol{\beta}_1'\left(X_1'M_1^*X_2\right) \\ &\quad \left[X_2'M_1^*X_2 + kI_{p_2}\right]^{-1}\left(X_2'M_1^*X_1\right)\boldsymbol{\beta}_1. \end{aligned}$$

We now assume that LSE of $\boldsymbol{\beta}$ exists, then we may also obtain ridge estimator using the marginal distribution theory as follows: since LS estimator $\tilde{\boldsymbol{\beta}}_n \sim N_p\left(\boldsymbol{\beta}, \sigma^2 \; C^{-1}\right)$, the marginal distribution of $\tilde{\boldsymbol{\beta}}_{1n}$ and of $\tilde{\boldsymbol{\beta}}_{2n}$ are $N_{p_1}\left(\boldsymbol{\beta}_1, \sigma^2 \; C_{11.2}^{-1}\right)$ and $N_{p_2}\left(\boldsymbol{\beta}_2, \sigma^2 \; C_{22.1}^{-1}\right)$, respectively.

We then define ridge regression estimator as

$$\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k) = \left[I_{p_1} + kC_{11.2}^{-1}\right]^{-1} \; \tilde{\boldsymbol{\beta}}_{1n},$$
$$\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k) = \left[I_{p_2} + kC_{22.1}^{-1}\right]^{-1}\tilde{\boldsymbol{\beta}}_{2n},$$

where $\tilde{\boldsymbol{\beta}}_{1n}$ and $\tilde{\boldsymbol{\beta}}_{2n}$ are the LS estimators and

$$C_{11.2} = X_1'X_1 - X_1'X_2 \; \left(X_2'X_2\right)^{-1}\left(X_2'X_1\right), \; C_{22.1} = X_2'X_2 - X_2'X_1 \; \left(X_1'X_1\right)^{-1}\left(X_1'X_2\right).$$

The bias and risk functions are given by

$$(1) \;\; \text{Bias}\left(\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k)\right) = -k\left[C_{11.2} + kI_{p_1}\right]^{-1} \; \boldsymbol{\beta}_1,$$
$$\text{Bias}\left(\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k)\right) = -k\left[C_{22.1} + I_{p_2}\right]^{-1} \; \boldsymbol{\beta}_2.$$

(2) $R\left(\tilde{\boldsymbol{\beta}}_{1n}^{ridge}(k);\boldsymbol{\beta}_1\right)= \sigma^2 tr\left\{[C_{11.2}+kI_{p_1}]^{-2}\ C_{11.2}\right\}+k^2\boldsymbol{\beta}_1'\ [C_{11.2}+kI_{p_1}]^{-2}\ \boldsymbol{\beta}_1,$

$R\left(\tilde{\boldsymbol{\beta}}_{2n}^{ridge}(k);\boldsymbol{\beta}_2\right)= \sigma^2 tr\left\{[C_{22.1}+kI_{p_2}]^{-2}\ C_{22.1}\right\}+k^2\boldsymbol{\beta}_2'\ [C_{22.1}+kI_{p_2}]^{-2}\ \boldsymbol{\beta}_2,$

respectively.

As a result, the quadratic error $\boldsymbol{D}_n^{*2}$ may be

$$\begin{aligned}\boldsymbol{D}_n^{*2} &= \left(Y-X\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right)'\ \left(Y-X\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right)\\ &= Y'Y-\left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right]'X'Y-k\left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right]'\ \left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right].\end{aligned}$$

The expression shows that if the total sum of squares less then the regression sum of squares for $\tilde{\boldsymbol{\beta}}_n^{ridge}(k)$ with modification depending on the squared length of $\tilde{\boldsymbol{\beta}}_n^{ridge}(k)$.

$$\begin{aligned}\boldsymbol{D}_n^{*2} &= \left(Y-X\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right)'\ \left(Y-X\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right)\\ &= Y'Y-\left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right]'X'Y-\left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right]'X'Y+\left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right]'\left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right],\end{aligned}$$

where

$$\begin{aligned}\left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right]'\left[I_n-\left(X'X+kI_p\right)^{-1}\right]X'Y &= -k\left[I_p+k\left(X'X\right)^{-1}\right]^{-1}\left(X'X\right)^{-1}X'Y\\ &= -k\left[X'X+kI_p\right]^{-1}X'Y=-k\tilde{\boldsymbol{\beta}}_n^{ridge}(k).\end{aligned}$$

Hence,

$$\boldsymbol{D}_n^{*2}=Y'Y-\left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right]'\ X'Y-k\left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right]'\ \left[\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right].$$

2.4 Ridge Trace and Estimation of k

Ridge regression estimator has been introduced to aid in tackling the issue of multicollinearity. It is difficult to untangle the relationships among the factors if one is confined to study the correlations out in the $X'X$ matrix. Thus, one needs some methods to have insight into the structure of the factor of space and sensitivity of the results to a particular

set of data at hand. Therefore, Hoerl and Kennard (1970) propose "ridge trace". Ridge trace is a graphical display of characterization of ridge regression. It is a two-dimensional plot of $\tilde{\boldsymbol{\beta}}_{jn}^{ridge}(k)$ against $k \geq 0$ and $\boldsymbol{D}_n^{*2} = \left(Y - X\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right)' \left(Y - X\tilde{\boldsymbol{\beta}}_n^{ridge}(k)\right)$ against $k \geq 0$. It serves to display the complex relationships that exists between nonorthogonal prediction vectors and the effect of these interrelationships on the estimation of $\boldsymbol{\beta}$. By computing $\tilde{\boldsymbol{\beta}}_{jn}^{ridge}(k)$ and $\boldsymbol{D}_n^{*2}$ for a set of values of k, such insight can be obtained. Figure 2.1 displays such a graph where we observe that reasonable coefficient stability is achieved in the range $0.4 < k_{opt} < 0.42$. The plots indicates roughly the estimate of k_{opt}.

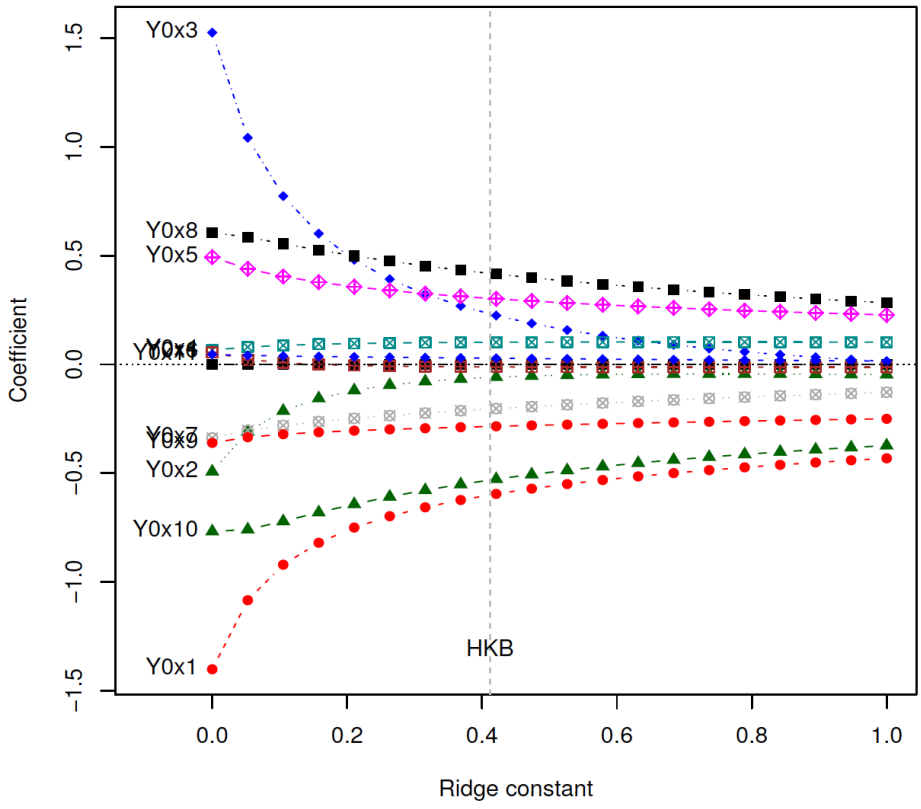

Figure 2.1: Ridge Trace: This figure is taken from Saleh et al. (2019).

Another method to estimate parameter k to give better estimate of $\boldsymbol{\beta}$ by damping the effect of the lower bound Ridge trace is a diagnostic test and guide to a better estimate of $\boldsymbol{\beta}$.
If k is chosen to be k_{opt}, then j^{th} components of $\tilde{\boldsymbol{\beta}}_{j,n}^{ridge}(k_{opt})$ is obtained to predict Y_j well. The generalized cross validation (GCV) is defined as the weighted average of predicted square errors.

$$\text{Var}(k) = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - x_j'\tilde{\boldsymbol{\beta}}_{jn}^{ridge}(k_{opt})\right)^2 w_j(k),$$

where $w_j(k) = (1 - a_{jj}(k)) \,/\, \left(1 - \frac{1}{n}\,\text{tr}\, A(k)\right)$, and $a_{jj}(k)$ is the j^{th} diagonal element of $A(k) = X\left(X'X + kI\right)^{-1}X'$. See Golub et al. (1979) for more details. Then,

$$k_{opt} = \arg\min_{k>0}\ \text{Var}(k).$$

The GCV theorem guarantees asymptotic efficiency of the GCV estimator for $p < n$ and also $p > n$.

Estimation of the parameter k is an important issue with ridge regression. There are many estimators depending on the model. To mention a few, we have Hoerl and Kennard (1970), Kibria (2003), Kibria and Banik (2016) and Kibria and Lukman (2020). Particularly, simple estimator by Lawless and Wang (1976) is given by

$$k^{LW} = \frac{\hat{\sigma}_n^2}{\tilde{\boldsymbol{\beta}}_n'\tilde{\boldsymbol{\beta}}_n}$$

for multiple linear regression model.

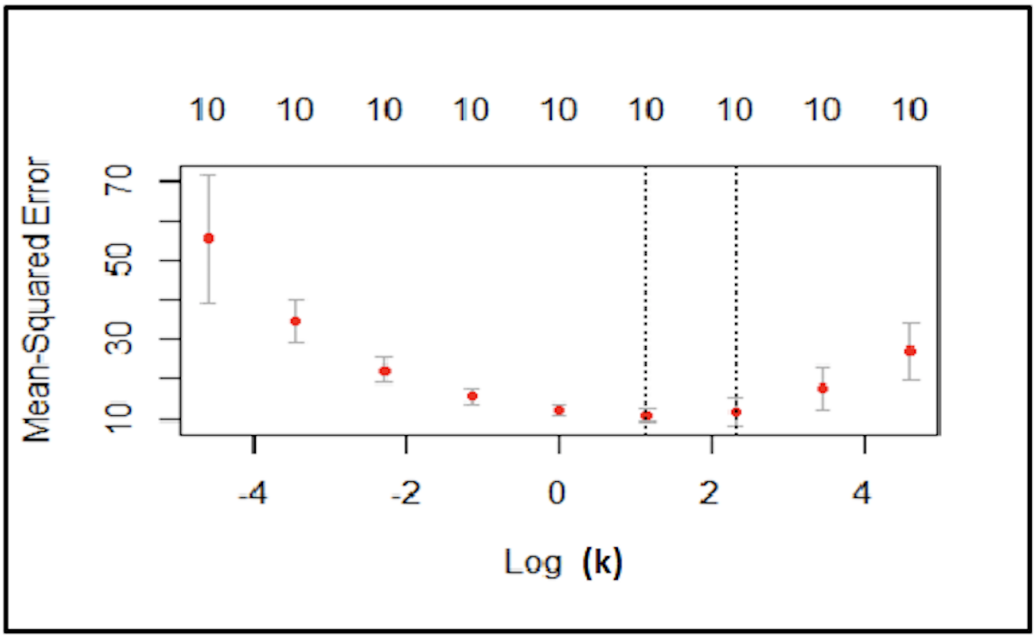

Figure 2.2: Plot of MSE vs. $log(k)$

From Figure 2.1, we can see that the value of the beta coefficients varies with the choice of the ridge constant $'k'$. For the different coefficients labeled using the colors, it can be seen that the values of these coefficients seem to stabilize and become more constant as we cross a certain threshold value of $'k'$, which happens to be about 0.42. Figure 2.2 is a related plot for simulated data that plots the MSE versus the logarithm of the ridge constant. As we see, the plot of the MSE, the values keep going down for increasing values of the ridge constant until they reach a minimum value, and then they rise back up again. The optimal values of log(k) lie between 1.4 and 2.2, which means that the optimal values of the ridge constant $'k'$ lie between 4.06 and 9.03. The values of the ridge constant changes as per the coefficients in the model and their significance. A future simulation study would include more experimentation with the data to study the behavior of the ridge parameters.

Chapter 3

LASSO

3.1 Some Subset Selectors

Many methods of subset selection, such as forward and backward selections among others have evolved around testing and estimating parameters. Now, it has settled down to preliminary test estimators with fixed critical value, proposed by Donoho and Johnstone (1994) as hard threshold estimators (HTE):

$$\hat{\boldsymbol{\beta}}_n^{PT}(\lambda) = \left(\tilde{\beta}_{jn} I\left(|\tilde{\beta}_{jn}| > \lambda\sigma\sqrt{c^{jj}}\right) | j = 1, \ldots, p\right)'. \tag{3.1}$$

We shall call it preliminary test subset selector (PTSS). However, this subset selector provides interpretable models that can be extremely variable because its discrete process-regressors are either dropped or retained from the model.

To remedy this problem with (3.1), Saleh (2006) proposed a continuous version of PTSS, starting with

$$\hat{\beta}_{jn}^{PT}(\lambda) = \tilde{\beta}_{jn} I\left(|\tilde{\beta}_{jn}| > \lambda\sigma\sqrt{c^{jj}}\right) = \tilde{\beta}_{jn} - \tilde{\beta}_{jn} I\left(|\tilde{\beta}_{jn}| \leq \lambda\sigma\sqrt{c^{jj}}\right). \tag{3.2}$$

Rewrite (3.2), we replace $I\left(|\tilde{\beta}_{jn}| \leq \lambda\sigma\sqrt{c^{jj}}\right)$ in the right hand-side by $\frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}$, $j = (1, \ldots, p)$. This yields the following Saleh-type selector (S)

$$\hat{\boldsymbol{\beta}}_n^{S}(\lambda) = \left(\tilde{\beta}_{jn}\left(1 - \frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right) | j = 1, \ldots, p\right)'. \tag{3.3}$$

This selector has the inherent problem of changing signs. Breiman (1996) obtained the non-negative garotte by minimizing

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}d_j\tilde{\beta}_{jn}x_{ij}\right)^2 \text{ subject to } d_j\geq 0 \text{ and } \sum_{j=1}^{p}d_j\leq t,\ t>0.$$

This means that we multiply $\tilde{\beta}_{jn}$ by a shrinkage factor, d_j. The result becomes

$$\hat{\boldsymbol{\beta}}_n^{(G)}(\lambda)=\left(\tilde{\beta}_{jn}\left(1-\frac{\lambda\sigma^2c^{jj}}{\hat{\beta}_{jn}^2}\right)|j=1,\ldots,p\right)'.$$

This selector does not change sign, and has its own problem to resolve. Let us look at the Saleh-type selector given by (3.3) and modify to make it positive-rule Saleh-type selector (PRSS) as follows

$$\hat{\boldsymbol{\beta}}_n^{S+}(\lambda)=\left(\tilde{\beta}_{jn}\left(1-\frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right)I\left(|\tilde{\beta}_{jn}|>\lambda\sigma\sqrt{c^{jj}}\right)|j=1,\ldots,p\right)'. \tag{3.4}$$

The formula (3.4) is called LASSO (Least Absolute Shrinkage and Selection Operator) proposed by Tibshirani (1996), which has gone viral in the literature. Mathematically, it is derived as shown in (3.5) and (3.6) below

$$\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda)=\arg\min_{\boldsymbol{\beta}\in R^p}\{\|Y-X\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}|,\ |\boldsymbol{\beta}|=\sum_{j=1}^{p}\boldsymbol{\beta}_j\} \tag{3.5}$$

$$=\left(\tilde{\beta}_{jn}\left(1-\frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right)I\left(|\tilde{\beta}_{jn}|>\lambda\sigma\sqrt{c^{jj}}\right)|j=1,\ldots,p\right)'. \tag{3.6}$$

The main advantage of this selector is that it shrinks some coefficients and sets others to *zero*. Hence, it retains good properties of ridge regression and subset selection procedures. We may notice that core statistics for subset selections are

$$\hat{\beta}_{jn}^{PT}(\lambda)=\sigma\sqrt{c^{jj}}\,\mathrm{sgn}(Z_{jn})|Z_{jn}|\left(1-I(|Z_{jn}|<\lambda)\right);\ Z_{jn}=\frac{\tilde{\beta}_{jn}}{\sigma\sqrt{c^{jj}}},$$

$$\hat{\beta}_{jn}^{S}(\lambda)=\sigma\sqrt{c^{jj}}\,\mathrm{sgn}(Z_{jn})|Z_{jn}|\left(1-\frac{\lambda}{|Z_{jn}|}\right),$$

$$\hat{\beta}_{jn}^{S+}(\lambda)=\sigma\sqrt{c^{jj}}\,\mathrm{sgn}(Z_{jn})|Z_{jn}|\left(1-\frac{\lambda}{|Z_{jn}|}\right)I\left(|Z_{jn}|>\lambda\right).$$

3.2 Risks of the Selectors

In this section, we present the L_2-risk expressions.
(a) First, we consider the PTSS as

$$
\begin{aligned}
&R\left(\hat{\boldsymbol{\beta}}_n^{PT}(\lambda);\boldsymbol{\beta}\right) = \sum_{j=1}^{p} \mathrm{E}\left(\hat{\beta}_{jn}^{PT}(\lambda) - \beta_j\right)^2 \\
&= \sum_{j=1}^{p} \mathrm{E}\left[\left(\tilde{\beta}_{jn} - \beta_j\right) - \tilde{\beta}_{jn}\, I\left(|\tilde{\beta}_{jn}| < \lambda\sigma\sqrt{c^{jj}}\right)\right]^2 \\
&= \sum_{j=1}^{p} \mathrm{E}\left[\left(\tilde{\beta}_{jn} - \beta_j\right)^2 + \tilde{\beta}_{jn}^2\, I\left(|\tilde{\beta}_{jn}| < \lambda\sigma\sqrt{c^{jj}}\right) - 2\left(\tilde{\beta}_{jn} - \beta_j\right)\tilde{\beta}_{jn}\right. \\
&\left. I\left(|\tilde{\beta}_{jn}| < \lambda\sigma\sqrt{c^{jj}}\right)\right] \\
&= \sigma^2 \sum_{j=1}^{p} \left[c^{jj} - \left(c^{jj} H_3\left(\lambda^2, \Delta_j^2\right) + \Delta_j^2 H_5\left(\lambda^2, \Delta_j^2\right)\right) + 2\beta_j^2 H_3\left(\lambda^2, \Delta_j^2\right)\right] \\
&;\ \Delta_j = \frac{\beta_{jn}}{\sigma\sqrt{c^{jj}}} \\
&= \sigma^2 \sum_{j=1}^{p} \left[c^{jj}\left(1 - H_3\left(\lambda^2, \Delta_j^2\right)\right) + \Delta_j^2\left(2H_3\left(\lambda^2, \Delta_j^2\right) - H_5\left(\lambda^2, \Delta_j^2\right)\right)\right],
\end{aligned}
$$

where $H_\gamma(.;\Delta^2)$ is the cdf of a non-central chi-square distribution with γ d.f. and non-central parameter Δ^2. It may be seen that the lower bound is given by

$$
R\left(\hat{\boldsymbol{\beta}}_n^{PT}(\lambda);\boldsymbol{\beta}\right) \geq \sigma^2\left(1 - H_3(\lambda^2;0)\right)\mathrm{tr}\, C^{-1}.
$$

Here, we have used the following results from Saleh (2006)

$$
\begin{aligned}
&\mathrm{E}\left[\tilde{\beta}_{jn} I(|\tilde{\beta}_{jn}| < \lambda)\right] = \beta_j H_3(\lambda^2;\Delta_j^2), \\
&\mathrm{E}\left[\tilde{\beta}_{jn}^2 I(|\tilde{\beta}_{jn}| < \lambda)\right] = \sigma^2\left[c^{jj} H_3(\lambda^2;\Delta_j^2) + \Delta_j^2 H_5(\lambda^2;\Delta_j^2)\right].
\end{aligned}
$$

(b) Now, we consider the Saleh-type selector as follows

$$
\begin{aligned}
R\left(\hat{\boldsymbol{\beta}}_n^S(\lambda);\boldsymbol{\beta}\right) &= \sum_{j=1}^{p} \mathrm{E}\left(\hat{\beta}_{jn}^S(\lambda)-\beta_j\right)^2 \\
&= \sum_{j=1}^{p} \mathrm{E}\left[\left(\hat{\beta}_{jn}^S(\lambda)-\beta_j\right)-\lambda\sigma\sqrt{c^{jj}}sgn\left(\tilde{\beta}_{jn}\right)\right]^2 \\
&= \sum_{j=1}^{p}\left[\mathrm{E}\left(\tilde{\beta}_{jn}-\beta\right)^2+\lambda^2\sigma^2c^{jj}-2\lambda\sigma\sqrt{c^{jj}}\mathrm{E}\left\{\tfrac{\left(\tilde{\beta}_{jn}-\beta_j\right)Z_{jn}}{|Z_{jn}|}\right\}\right] \\
&= \sigma^2\left[\sum_{j=1}^{p}c^{jj}+\lambda^2\sum_{j=1}^{p}c^{jj}-2\lambda\sum_{j=1}^{p}c^{jj}\left(\mathrm{E}|Z_{jn}|-\Delta_j\mathrm{E}\frac{Z_{jn}}{|Z_{jn}|}\right)\right] \\
&= \sigma^2\left[\mathrm{tr}\,C^{-1}\left(1+\lambda^2\right)-2\lambda\sum_{j=1}^{p}c^{jj}\sqrt{\frac{2}{\pi}}e^{\frac{-\Delta_j^2}{2}}\right],
\end{aligned}
$$

where $Z_{jn}=\frac{\hat{\beta}_{jn}}{\sigma\sqrt{c^{jj}}},\ j=1,\ldots,p.$

Differentiating with respect to λ we have the optimum value for λ given by

$$
\lambda_{opt}=\frac{\sum_{j=1}^{p}c^{jj}\sqrt{\frac{2}{\pi}}\ e^{-\frac{\Delta^2}{\pi}}}{\mathrm{tr}\,C^{-1}}=\sqrt{\frac{2}{\pi}}\ \text{when}\ \Delta_j=0\ \forall j.
$$

Therefore,

$$
\begin{aligned}
R\left(\hat{\boldsymbol{\beta}}_n^S(\lambda);\boldsymbol{\beta}\right) &= \sigma^2\left[\left(1+\frac{2}{\pi}\right)\mathrm{tr}\,C^{-1}-\frac{4}{\pi}\ \sum_{j=1}^{p}c^{jj}c^{-\frac{\Delta_j^2}{2}}\right] \\
&= \sigma^2\left[\mathrm{tr}\,C^{-1}-\frac{2}{\pi}\ \left(\sum_{j=1}^{p}c^{jj}\left(2\ e^{-\frac{\Delta_j^2}{2}}-1\right)\right)\right].
\end{aligned}
$$

(c) Positive-rule Saleh-type estimator or LASSO:
In this section, we consider L_2-risk of the positive-rule Saleh-estimator

$$
\begin{aligned}
&R\left(\hat{\boldsymbol{\beta}}_n^{S+}(\lambda);\boldsymbol{\beta}\right) = \mathrm{E}\left(\hat{\boldsymbol{\beta}}_n^{S+}(\lambda)-\boldsymbol{\beta}\right)'\left(\hat{\boldsymbol{\beta}}_n^{S+}(\lambda)-\boldsymbol{\beta}\right)\\
&= \sum_{j=1}^{p}\mathrm{E}\left[\left(\tilde{\beta}_{jn}\left(1-\frac{\lambda^*}{|\tilde{\beta}_{jn}|}\right)-\beta_j\right)-\hat{\beta}_{jn}\left(1-\frac{\lambda^*}{|\tilde{\beta}_{jn}|}\right)I\left(|\tilde{\beta}_{jn}|<\lambda^*\right)\right]^2\\
&= \sum_{j=1}^{p}\mathrm{E}\left[\left\{\tilde{\beta}_{jn}\left(1-\frac{\lambda^*}{|\tilde{\beta}_{jn}|}\right)-\beta_j\right\}^2-\tilde{\beta}_{jn}^2\left(1-\frac{\lambda^*}{|\tilde{\beta}_{jn}|}\right)I\left(|\tilde{\beta}_{jn}|<\lambda^*\right)\right.\\
&\left.+2\beta_j\tilde{\beta}_{jn}\left(1-\frac{\lambda^*}{|\tilde{\beta}_{jn}|}\right)I\left(|\tilde{\beta}_{jn}|<\lambda^*\right)\right]\\
&= R\left(\hat{\boldsymbol{\beta}}_n^{S}(\lambda);\boldsymbol{\beta}\right)-\sigma^2\sum_{j=1}^{p}\left\{c^{jj}\mathrm{E}\left[Z_{jn}^2\left(1-\frac{\lambda}{|Z_{jn}|}\right)^2 I\left(|Z_{jn}|<\lambda\right)\right]\right.\\
&\left.+2\Delta_j\mathrm{E}\left[Z_{jn}\left(1-\frac{\lambda}{|Z_{jn}|}\right)I\left(|Z_{jn}|<\lambda\right)\right]\right\}\geq 0\ ,\\
&= \sigma^2\sum_{j=1}^{p}\left[c^{jj}\left\{\left(1-H_3\left(\lambda_j^2,\Delta_j^2\right)\right)+\Delta_j^2\left[2H_3\left(\lambda_j^2,\Delta_j^2\right)\right.\right.\right.\\
&\left.\left.\left.-H_5\left(\lambda_j^2,\Delta_j^2\right)\right]+\lambda^2\left(1-H_1\left(\lambda_j^2,\Delta_j^2\right)\right)-2\lambda\left(\phi\left(\lambda-\Delta_j\right)+\phi\left(\lambda+\Delta_j\right)\right)\right\}\right].
\end{aligned}
$$

3.3 Alternative Derivation of Bias and Risk expression of LASSO

In this section, we derive the bias and risk expressions of LASSO. We state the following theorem for them.

Theorem 3.3.1. *If $Z\sim N(\Delta,1)$, then the bias and risk expressions are given by*

$$
\begin{aligned}
\mathrm{Bias}\left(Z^{LASSO}(\lambda)\right) &= \mathrm{E}\left[Z^{LASSO}(\lambda)\right]-\Delta\\
&= -\{[\Phi\left(\lambda-\Delta\right)-\Phi\left(-\lambda-\Delta\right)]-[\phi\left(\lambda-\Delta\right)-\phi\left(\lambda+\Delta\right)]\\
&\quad -\lambda\left[\Phi\left(\lambda-\Delta\right)-\Phi\left(\lambda+\Delta\right)\right]\},\\
R\left(Z^{LASSO}(\lambda);Z\right) &= \mathrm{E}\left[Z^{LASSO}(\lambda)-\Delta\right]^2\\
&= 1+\lambda^2+\left(\Delta^2-\lambda^2-1\right)\left[\Phi\left(\lambda-\Delta\right)-\Phi\left(-\lambda-\Delta\right)\right]\\
&\quad -\left[\left(\lambda-\Delta\right)\phi\left(\lambda+\Delta\right)+\left(\lambda+\Delta\right)\phi\left(\lambda-\Delta\right)\right],
\end{aligned}
$$

where $Z^{LASSO}(\lambda) = \text{sgn}(Z)\left(|Z| - \lambda\right)^{+} = Z\left(1 - \frac{\lambda}{|Z|}\right) I\left(|Z| > \lambda\right)$.

For the proof, we need the following lemma proposed by Saleh et al. (2019).

Lemma 3.3.2. *If* $Z \sim N(\Delta, 1)$ *then:*

(i) $\text{E}[|Z|] = \Delta[2\Phi(\Delta) - 1] + \sqrt{\frac{2}{\pi}}\ \exp\{-\frac{\Delta^2}{2}\}$,

(ii) $\text{E}[I\left(|Z| < \lambda\right)] = [\Phi\left(\lambda - \Delta\right) - \Phi\ \ (-\lambda - \Delta)]$,

(iii) $\text{E}[ZI\left(|Z| < \lambda\right)] = \Delta\ [\Phi\left(\lambda - \Delta\right) - \Phi\left(-\lambda - \Delta\right)] - [\phi\left(\lambda - \Delta\right) - \phi\left(\lambda + \Delta\right)]$,

(iv) $\text{E}[sgn(Z)] = 2\Phi(\Delta) - 1$,

(v) $\text{E}[sgn(Z)\ I\left(|Z| < \lambda\right)] = [\Phi\left(\lambda - \Delta\right)\ - \Phi\left(\lambda + \Delta\right)] + [2\ \Phi(\Delta)\ - 1]$,

(vi) $\text{E}[|Z|\ I\left(|Z| < \lambda\right)\ ] = [\phi(\Delta) - \phi\left(\lambda + \Delta\right)\ ] + [\phi(\Delta) - \phi\left(\lambda - \Delta\right)\ ]$
$- \Delta[\Phi\left(\lambda + \Delta\right)\ - \Phi\left(\lambda - \Delta\right)\ + 1] + 2\ \Delta\Phi(\Delta)\ ,$

(vii) $\text{E}[Z^2\ I\left(|Z| < \lambda\right)] = 2\ \Delta\ [\phi\left(\lambda + \Delta\right)\ - \phi\left(\lambda - \Delta\right)\ ] - [(\lambda - \Delta)\ \phi\left(\lambda - \Delta\right)$
$+ (\lambda + \Delta)\ \phi\left(\lambda + \Delta\right)] + (\Delta^2\ + 1)\ [\Phi\left(\lambda - \Delta\right)\ + \Phi\left(\lambda + \Delta\right) - 1]$,

where Φ *is the cdf of N(0,1) and* $H_\gamma(., \Delta^2)$ *is the cdf of a non-central chi-square distribution with* γ *degree of freedom and non-central parameter* $\frac{\Delta^2}{2}$.

Proof. See Saleh et al. (2019).

□

Theorem 3.3.3. *The bias of LASSO estimator and the risk of LASSO estimator are given by*

$$\begin{aligned}\text{Bias}\left(\hat{\beta}_{jn}^{LASSO}(\lambda)\right) = \sigma\sqrt{c^{jj}}\ \{\lambda\left[\Phi(\lambda - \Delta) - \Phi(\lambda + \Delta)\right] - \Delta\left[\Phi(\lambda - \Delta)\right. \\ \left. -\Phi(-\lambda - \Delta)\right] + \left[\phi(\lambda - \Delta) - \phi(\lambda + \Delta)\right]\}\,,\end{aligned}$$

$$R\left(\hat{\beta}_{jn}^{LASSO}(\lambda);\beta_j\right) = \sigma\sqrt{c^{jj}}\left\{\left(\mathrm{E}\left[Z_{jn}^2\right] + \lambda^2 - 2\ \lambda\ \mathrm{E}\left[|Z_{jn}|\right]\right) - \left(\mathrm{E}\left[Z_{jn}^2\right.\right.\right.$$
$$\left.\left.\left. I\left(|Z_{jn}| < \lambda\right)\right] + \lambda^2\ \mathrm{E}\left[I\left(|Z_{jn}| < \lambda\right)\right] - 2\ \lambda\ \mathrm{E}\left[|Z_{jn}| I\left(|Z_{jn}| < \lambda\right)\right]\right)$$
$$+2\ \Delta_j\left(\mathrm{E}\left[Z_{jn}\right] - \lambda\ \mathrm{E}\left[\mathrm{sgn}(Z_{jn})\right]\right) + \Delta_j^2 - 2\ \Delta_j\left(\mathrm{E}\left[Z_{jn} I(|Z_{jn}| < \lambda)\right]\right.$$
$$\left.-\lambda\ \mathrm{E}\left[\mathrm{sgn}(Z_{jn}) I(|Z_{jn}| < \lambda)\right]\right)\}.$$

Proof. By substituting Lemma (3.3.2) (iii) and (v), we obtain the following:

$$\begin{aligned}
&\mathrm{Bias}\left(\hat{\beta}_{jn}^{LASSO}(\lambda)\right) = \mathrm{E}\left[\hat{\beta}_{jn}^{LASSO}(\lambda)\right] - \beta_j \\
&= \mathrm{E}\left[\tilde{\beta}_{jn}\left(1 - \frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right) I\left(|\tilde{\beta}_{jn}| > \lambda\sigma\sqrt{c^{jj}}\right) - \beta_j\right] \\
&= \mathrm{E}\left[\left(\tilde{\beta}_{jn} - \beta_j\right) - \mathrm{sgn}(\tilde{\beta}_{jn})\lambda\sigma\sqrt{c^{jj}}\right. \\
&\left.-\tilde{\beta}_{jn}\left(1 - \frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right) I\left(|\tilde{\beta}_{jn}| < \lambda\sigma\sqrt{c^{jj}}\right)\right] \\
&= \sigma\sqrt{c^{jj}}\ \mathrm{E}\left\{-\lambda\,\mathrm{sgn}(Z_{jn}) - \mathrm{E}\left[Z_{jn}\left(1 - \frac{\lambda}{|Z_{jn}|}\right) I\left(|Z_{jn}| \le \lambda\right)\right]\right\} \\
&= -\sigma\sqrt{c^{jj}}\left\{\mathrm{E}\left[Z_{jn}\ I(|Z_{jn}| < \lambda)\right] + \lambda\ \mathrm{E}\left[\mathrm{sgn}(Z_{jn})\right]\right. \\
&\left.-\lambda\ \mathrm{E}\left[\mathrm{sgn}(Z_{jn}) I(|Z_{jn}| < \lambda)\right]\right\} \\
&= \sigma\sqrt{c^{jj}}\left\{\Delta\left[\Phi(\lambda - \Delta) - \Phi(-\lambda - \Delta)\right] - \left[\phi(\lambda - \Delta) - \phi(\lambda + \Delta)\right]\right. \\
&\left.+\lambda(2\ \Phi(\Delta) - 1) - \lambda\left[\Phi(\lambda - \Delta) - \Phi(\lambda + \Delta)\right] - (2\ \Phi(\Delta - 1))\right\} \\
&= \sigma\sqrt{c^{jj}}\left\{\lambda\left[\Phi(\lambda - \Delta) - \Phi(\lambda + \Delta)\right] - \Delta\left[\Phi(\lambda - \Delta) - \Phi(-\lambda - \Delta)\right]\right. \\
&\left.+\left[\phi(\lambda - \Delta) - \phi(\lambda + \Delta)\right]\right\}.
\end{aligned}$$

Now, by using results from Lemma (3.3.2), we obtain the expression of risk as follows

$$\begin{aligned}
&R\left(\hat{\beta}_{jn}^{LASSO}(\lambda);\beta_j\right) = \mathrm{E}\left(\hat{\beta}_{jn}^{LASSO}(\lambda) - \beta_j\right)^2 \\
&= \mathrm{E}\left[\tilde{\beta}_{jn}\left(1 - \frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right) I\left(|\tilde{\beta}_{jn}| > \lambda\sigma\sqrt{c^{jj}}\right) - \beta_j\right]^2
\end{aligned}$$

$$= \sigma\sqrt{c^{jj}}\ \mathrm{E}\left[Z_{jn}\left(1-\frac{\lambda}{|Z_{jn}|}\right) - Z_{jn}\left(1-\frac{\lambda}{|Z_{jn}|}\right) I\left(|Z_{jn}|<\lambda\right) - \Delta_j\right]^2$$

$$= \sigma\sqrt{c^{jj}}\ \mathrm{E}\left[Z_{jn}^2\left(1-\frac{\lambda}{|Z_{jn}|}\right)^2 - Z_{jn}^2\left(1-\frac{\lambda}{|Z_{jn}|}\right)^2 I\left(|Z_{jn}|<\lambda\right) + \Delta_j^2\right.$$
$$\left. +2\ \Delta_j Z_{jn}\left(1-\frac{\lambda}{|Z_{jn}|}\right) - 2\ \Delta_j Z_{jn}\left(1-\frac{\lambda}{|Z_{jn}|}\right) I\left(|Z_{jn}<\lambda\right)\right]$$

$$= \sigma\sqrt{c^{jj}}\left\{\left(\mathrm{E}\left[Z_{jn}^2\right] + \lambda^2 - 2\ \lambda\ \mathrm{E}\left[|Z_{jn}|\right]\right) - \left(\mathrm{E}\left[Z_{jn}^2 I\left(|Z_{jn}|<\lambda\right)\right]\right.\right.$$
$$+\lambda^2\ \mathrm{E}\left[I\left(|Z_{jn}|<\lambda\right)\right] - 2\ \lambda\ \mathrm{E}\left[|Z_{jn}| I\left(|Z_{jn}|<\lambda\right)\right]\Big)$$
$$+2\ \Delta_j\left(\mathrm{E}\left[Z_{jn}\right] - \lambda\ \mathrm{E}\left[\mathrm{sgn}(Z_{jn})\right]\right) + \Delta_j^2$$
$$\left.-2\ \Delta_j\left(\mathrm{E}\left[Z_{jn} I(|Z_{jn}|<\lambda)\right] - \lambda\ \mathrm{E}\left[\mathrm{sgn}(Z_{jn}) I(|Z_{jn}|<\lambda)\right]\right)\right\}.$$

□

Theorem 3.3.4. *Risk of $\hat{\beta}_{jn}^{LASSO}(\lambda)$ is less than that of $\hat{\beta}_{jn}^{S}(\lambda)$.*

Proof.

$$R\left(\hat{\beta}_{jn}^{LASSO}(\lambda);\beta_j\right) = \mathrm{E}\left[\hat{\beta}_{jn}^{LASSO}(\lambda) - \beta_j\right]^2$$

$$= \mathrm{E}\left[\tilde{\beta}_{jn}\left(1-\frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right) I\left(|\tilde{\beta}_{jn}| > \lambda\sigma\sqrt{c^{jj}} - \beta_j\right)\right]^2$$

$$= \mathrm{E}\left[\left(\tilde{\beta}_{jn}\left(1-\frac{\lambda}{|\tilde{\beta}_{jn}|}\right) - \tilde{\beta}_{jn}\right) - \tilde{\beta}_{jn}\left(1-\frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right)\right.$$
$$\left. I\left(|\tilde{\beta}_{jn}| < \lambda\sigma\sqrt{c^{jj}} - \beta_j\right)\right]^2$$

$$= \mathrm{E}\left[\left(\tilde{\beta}_{jn}\left(1-\frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right) - \beta_j\right)^2 - \tilde{\beta}_{jn}^2\left(1-\frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right)^2\right.$$
$$\left. I\left(|\tilde{\beta}_{jn}| < \lambda\sigma\sqrt{c^{jj}}\right) + 2\ \beta_j\left(\tilde{\beta}_{jn} - \lambda\sigma\sqrt{c^{jj}}\,\mathrm{sgn}(\tilde{\beta}_{jn})\right) I\left(|\tilde{\beta}_{jn}| < \lambda\right)\right]$$

$$= R\left(\hat{\beta}_{jn}^{S}(\lambda);\beta_j\right) - \mathrm{E}\left[\tilde{\beta}_{jn}^2\left(1-\frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right)^2 I\left(|\tilde{\beta}_{jn}| < \lambda\sigma\sqrt{c^{jj}}\right)\right.$$
$$\left. +2\ \beta_j\left(\tilde{\beta}_{jn} - \lambda\sigma\sqrt{c^{jj}}\,\mathrm{sgn}\,\tilde{\beta}_{jn}\right) I\left(|\tilde{\beta}_{jn}| < \lambda\sigma\sqrt{c^{jj}}\right)\right] \geq 0\ \forall\ \beta_{jn} \in R.$$

Hence, $R\left(\hat{\beta}_{jn}^{LASSO}(\lambda);\beta_j\right) \leq R\left(\hat{\beta}_{jn}^{S}(\lambda);\beta_j\right)\ \forall\ \beta_j \in R.$

□

Table 3.1: Efficiency of restricted, ridge, PTE, Saleh-type, and LASSO estimators for location parameter.

Δ^2	$\hat{\theta}_n$	$\tilde{\theta}_n^{ridge}$	$\hat{\theta}_n^{PT}$	$\hat{\theta}_n^{S}$	$\hat{\theta}_n^{S+}$
0	∞	∞	1.411	2.752	4.303
0.01	100.000	101.000	1.399	2.705	4.199
0.02	50.000	51.000	1.387	2.659	4.101
0.03	33.333	34.333	1.375	2.615	4.007
0.04	25.000	26.000	1.364	2.573	3.918
0.05	20.000	21.000	1.353	2.533	3.833
0.1	10.000	11.000	1.301	2.350	3.463
0.2	5.000	6.000	1.213	2.064	2.916
0.3	3.333	4.333	1.141	1.849	2.532
0.4	2.500	3.500	1.082	1.683	2.248
0.5	2.000	3.000	1.031	1.550	2.029
1	1.000	2.000	0.869	1.157	1.417
2ln(2)	0.721	1.721	0.801	1.000	1.187
2	0.500	1.500	0.743	0.856	0.980
3	0.333	1.333	0.713	0.739	0.813
4	0.250	1.250	0.720	0.683	0.730
5	0.200	1.200	0.746	0.653	0.685
10	0.100	1.100	0.908	0.614	0.619
20	0.050	1.050	0.996	0.611	0.611
30	0.033	1.033	1.000	0.611	0.611
40	0.025	1.025	1.000	0.611	0.611
50	0.020	1.020	1.000	0.611	0.611

From the Table 3.1, we observe that ridge estimator $\hat{\theta}_n^{ridge}$ dominates all others in the table and $\hat{\theta}_n^{S+}$ is the next best estimator dominating $\hat{\theta}_n^{S}$ uniformly. Now, look at $\hat{\theta}_n$ and $\hat{\theta}_n^{PT}$: $\hat{\theta}_n$ dominates $\hat{\theta}_n^{PT}$, $\hat{\theta}_n^{S}$ and $\hat{\theta}_n^{S+}$ in a limited interval. Same holds for $\hat{\theta}_n^{PT}$. Thus, one may prefer ridge and LASSO depending on what problem one is solving.

With the Table 3.1, it can be seen that the next estimator is the positive rule Stein-Saleh type estimator (LASSO) dominates only Saleh-type estimator. Neither PTE nor any other estimator except ridge dominates the others uniformly.

Chapter 4

Elastic net

In this chapter, we consider the elastic net, a regularization and variable selection method introduced by Zou and Hastie (2005). Elastic net is a combination of L_1 an L_2 penalty function to obtain LASSO and ridge regression estimators respectively. We observed in chapter 2 and 3 that ridge regression is a shrinkage estimator that encourages the variables to be small but not be to 0, while LASSO sets some of the variables to 0 and selects only one variable in a group of correlated variables. If $p > n$, LASSO selects at most n variables which is a limiting feature of LASSO. Further, for $n > p$, performance of LASSO is dominated by the ridge regression estimator.

To elevate these problems, one may use elastic net. There are two kinds of elastic net, namely, naive elastic net and elastic net. In this chapter, we study the performance properties of elastic net estimator. We begin with the location model in the next section.

4.1 Elastic net Subset Selector

In this section, we derive and study the naive elastic net (nEnet) for the location and regression model and introduce the elastic net (Enet).

4.1.1 Location Model

First, we consider the location model

$$Y = \theta 1_n + \epsilon_n, \quad \epsilon_n \sim N_n\left(0, \sigma^2 I_n\right). \tag{4.1}$$

Then, the ridge regression estimator of (4.1) form is given by

$$\begin{aligned}\hat{\theta}_n^{ridge}(\lambda) &= \arg\min_{\theta}\{(Y-\theta 1_n)'(Y-\theta 1_n)+n\theta^2\} = \frac{\bar{Y}}{1+\lambda} \\ &= \frac{\sigma}{\sqrt{n}}\,\frac{Z}{1+\lambda}\;; \quad Z=\frac{\sqrt{n}\bar{Y}}{\sigma},\end{aligned} \tag{4.2}$$

and the LASSO of (4.1) form is written as

$$\begin{aligned}\hat{\theta}_n^{LASSO}(\lambda) &= \arg\min_{\theta}\{(Y-\theta 1_n)'(Y-\theta 1_n)+\lambda\sigma\sqrt{n}|\theta|\} \\ &= \frac{\sigma}{\sqrt{n}}\operatorname{sgn} Z\,(|Z|-\lambda)^+.\end{aligned} \tag{4.3}$$

Now, we consider the elastic net estimator

$$\begin{aligned}\hat{\theta}_n^{nEnet}(\lambda,\alpha) &= \arg\min_{\theta}\left[(Y-\theta 1_n)'(Y-\theta 1_n)+\lambda\left(\sqrt{n}\alpha\sigma|\theta|+\frac{n}{2}(1-\alpha)\theta^2\right)\right] \\ &= \frac{\sigma}{\sqrt{n}}\,\frac{sgn(Z)\,(|Z|-\lambda\alpha)^+}{[1+\lambda(1-\alpha)]}.\end{aligned} \tag{4.4}$$

Notice that (4.4) it is a ridge regression type estimator using LASSO. It must be better than LASSO depending on the values of (λ,α). It is uniformly better than the ridge regression obtained for $\alpha=0$. See the Table 4.1.

Note that from the formula (4.4), we can obtain the forms (4.2) and (4.3), respectively as follows

$$\begin{aligned}\text{If } \alpha=0, \;&= \frac{\sigma}{\sqrt{n}}\,\frac{Z}{[1+\lambda]} = \hat{\theta}_n^{ridge}(\lambda), \\ \text{if } \alpha=1, \;&= \frac{\sigma}{\sqrt{n}}\operatorname{sgn} Z\,(|Z|-\lambda)^+ = \hat{\theta}_n^{LASSO}(\lambda).\end{aligned}$$

The bias and risk of this estimator are given by

$$\begin{aligned}\text{Bias}\left(\hat{\theta}_n^{n\ \text{Enet}}(\lambda,\alpha)\right) &= -\frac{\sigma}{\sqrt{n}}[1+\lambda(1-\alpha)]^{-1}\left[\Delta H_3\left(\lambda^2\alpha^2;\Delta^2\right)\right. \\ &\quad \left.-\lambda\alpha(\Phi(\lambda\alpha-\Delta)-\Phi(\lambda\alpha+\Delta))\right];\Delta^2=\frac{n\theta^2}{\sigma^2},\end{aligned}$$

$$R\left(\hat{\theta}_n^{nEnet}(\lambda,\alpha);\theta\right) = \frac{\sigma^2}{n}\left[1+\lambda(1-\alpha)\right]^{-2}\left[\frac{n}{\sigma^2}R\left(\hat{\theta}_n^{S+}(\lambda,\alpha);\theta\right) + \lambda^2(1-\alpha)^2\right.$$
$$\left.\Delta^2 + \frac{2}{n}\lambda(1-\alpha)\Delta\{\Delta H_3\left(\lambda^2\alpha^2;\Delta^2\right) - \lambda\alpha\left(\Phi\left(\lambda\alpha-\Delta\right) - \Phi\left(\lambda\alpha+\Delta\right)\right)\}\right].$$

The following table and graphs give the efficiency of nEnet and Enet estimators.

Table 4.1: Efficiency table for nEnet and Enet: The table is taken from the book: Rank-Based Methods for Shrinkage and Selection with Application to Machine Learning, Saleh et al. (2021).

$\alpha \mid \Delta^2$	Estimators	0	0.05	0.1	0.5	1	$\sqrt{2\ln(2)}$	2	5
0	nEnet	3.2324	3.2258	3.2063	2.6955	1.8452	1.3633	0.8810	0.1911
	sEnet	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.1	nEnet	3.3586	3.3397	3.3065	2.6851	1.8018	1.3362	0.8797	0.2041
	sEnet	1.1378	1.1337	1.1298	1.1025	1.0765	1.0609	1.0421	1.0037
0.2	nEnet	3.4870	3.4535	3.4045	2.6705	1.7602	1.3105	0.8788	0.2198
	sEnet	1.2991	1.2888	1.2789	1.2114	1.1498	1.1143	1.0728	0.9925
0.5	nEnet	3.8693	3.7733	3.6624	2.5954	1.6458	1.2416	0.8776	0.2910
	sEnet	1.9771	1.9244	1.8750	1.5684	1.3285	1.2056	1.0748	0.8660
0.8	nEnet	4.1882	3.9850	3.7827	2.4326	1.5294	1.1742	0.8680	0.4230
	sEnet	3.1148	2.9485	2.7966	1.9538	1.4136	1.1762	0.9517	0.6565
0.9	nEnet	4.2598	4.0041	3.7620	2.3405	1.4798	1.1448	0.8566	0.4779
	sEnet	3.6535	3.4207	3.2100	2.0858	1.4201	1.1446	0.8952	0.5857
1	nEnet	4.3030	3.9826	3.6948	2.2194	1.4173	1.1061	0.8363	0.5190
	sEnet	4.3030	3.9826	3.6948	2.2194	1.4173	1.1061	0.8363	0.5190

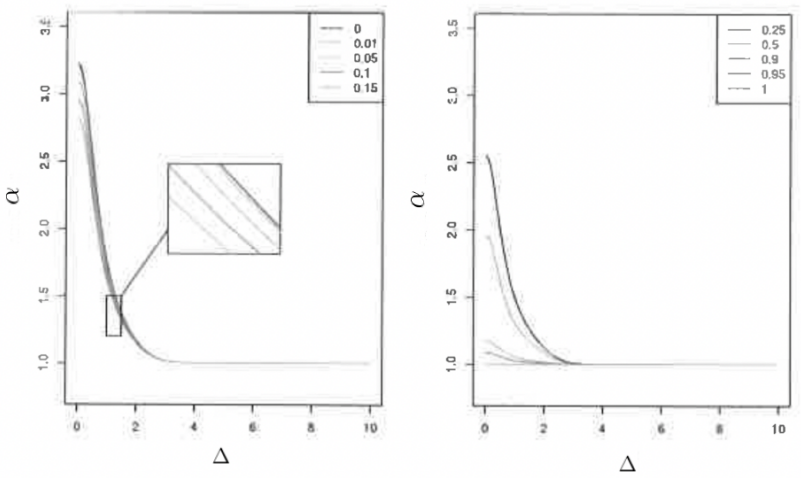

Figure 4.1: Graphs of Efficiencies: This graph obtains from book: Rank-Based Methods for Shrinkage and Selection with Application to Machine Learning, Saleh et al. (2021)

From the Table 4.1, we observe for both of Enet and nEnet that the efficiency decreases as λ-values increase for every α for nEnet. And for Enet the efficiency decreases as α increases for some λ. However, nEnet does better than Enet always. For $\alpha = 0$, nEnet become simple ridge estimator and for $\alpha = 1$, it becomes LASSO, with λ as tunning parameter. Note that for $\alpha = 1$ the penalty function is convex but not strictly convex. Naive elastic net can be looked at as a two-step procedure: a LASSO-type subset selection followed by ridge-type shrinkage. On the other hand, Enet is scaled version of nEnet, removing ridge-type shrinkage. Hence, all good properties of LASSO hold for Enet. From the Table 4.1, it is clear that nEnet outperforms Enet lagging behind, contrary to Zou and Hastie (2005).

4.2 Multiple regression Model and Elastic net

Consider The multiple regression model

$$Y_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N_p\left(0, \sigma^2 I_n\right).$$

We have two cases to consider : (i) $p < n$ and (ii) $p > n$. For the case $p < n$, we notice that ridge-type estimator dominate over LASSO-type estimator. This property continues to hold for multiple regression. However, if $p > n$, then one may consider Enet with some mathematical procedure. First, note that we have ridge regression estimator

$$\tilde{\boldsymbol{\beta}}_n^{ridge}(\lambda) = \left(X'X + \lambda^* I\right)^{-1} X'Y,\ \lambda^* = \lambda(1-\alpha).$$

Opposed to the LS estimator, $\tilde{\boldsymbol{\beta}}_n = \left(X'X\right)^{-1} X'Y$. For $p < n$, if $X'X$ is of full rank, we may write

$$\tilde{\boldsymbol{\beta}}_n^{ridge}(\lambda) = \left(I_p + \lambda^* C_n^{-1}\right)^{-1} \tilde{\boldsymbol{\beta}}_n, \quad C_n = \frac{1}{n} X'X.$$

Similarly, the LASSO is given by

$$\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda) = \left(\operatorname{sgn}(\tilde{\beta}_{jn}) \left(|\tilde{\beta}_{jn}| - \lambda\alpha\sigma\sqrt{c^{jj}}\right)^+ | j = 1, \ldots, p\right)'.$$

To obtain nEnet estimator of $\boldsymbol{\beta}$, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda) = \operatorname{argmin} \big\{ &(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}) \\ &+\lambda \left[\alpha\sigma \operatorname{diag}\left([c^{jj}]^{-\frac{1}{2}}|\beta_j|\ | j = 1, \ldots, p\right) + \frac{1}{2}(1-\alpha)\boldsymbol{\beta}'\boldsymbol{\beta}\right]\Big\}.\end{aligned}$$

Thus, we solve the following equation

$$\left[X'X + \lambda(1-\alpha)I_p\right]\boldsymbol{\beta} = X'Y - \lambda\alpha\sigma \operatorname{diag}\left([c^{jj}]^{\frac{1}{2}} \operatorname{sgn}(\tilde{\beta}_{jn}) | j = 1, \ldots, p\right)'.$$

To obtain

$$\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda, \alpha) = \left[I_p + \lambda(1-\alpha)C_n^{-1}\right]^{-1}\ \hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha),$$

where

$$\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha) = \left(\sigma\sqrt{c^{ii}}[1 + \lambda\,(1-\alpha)\,c^{jj}]^{-1} \operatorname{sgn}(Z_{jn})\,(|Z_{jn}| - \lambda\alpha)^+ | j = 1, \ldots, p\right)'.$$

Alternatively, we can define rescaled naive elastic net as Enet given by

$$\hat{\boldsymbol{\beta}}_n^{Enet}(\lambda, \alpha) = [1 + \lambda(1-\alpha)]\left[I_p + \lambda(1-\alpha)C_n^{-1}\right]\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha). \tag{4.5}$$

The rescaled naive elastic net given by (4.5) is called elastic net by Zou and Hastie(2005).

Using a family of diagonal linear projections, we obtain subset selection as

$$\left(\tilde{\boldsymbol{\beta}}'_{1n}, \mathbf{0}'\right)', \quad p_1 + p_2 = p.$$

As for nEnet, we obtain shrunken subset selector as

$$\begin{pmatrix} \left[I_{p_1} + \lambda^* C^{-1}_{n11.2}\right]^{-1} \tilde{\boldsymbol{\beta}}_{1n} \\ \mathbf{0} \end{pmatrix},$$

where $C_{n11.2} = C_{n11} - C_{n12}C^{-1}_{n22}C_{n21}$.

Recalling $C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$. Then, lower bound of the risk is given by

$$R^*\left(\boldsymbol{\beta}_n^{LASSO}(\lambda\alpha) : \boldsymbol{\beta}\right) \geq \sigma^2 \left[\text{tr}\left\{\left(C_{11.2} + \lambda(1-\alpha)I_{p_1}\right)^{-2} C_{11.2}\right\} + \Delta^2 Ch_{min}(C^{-1}_{22.1})\right]$$

where $\Delta^2 = \frac{1}{\sigma^2}\boldsymbol{\beta}'_2 C_{n22.1}\boldsymbol{\beta}_2$ and $Ch_{min}(A)$ is the smallest eigenvalue of A. This result is obtained from Donoho and Johnstone (1994). For the risk expression of nEnet, we obtain

$$R\left(\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha);\boldsymbol{\beta}\right) = \sigma^2\left[\text{tr}\left\{\left(C_{11.2} + \lambda(1-\alpha)I_{p_1}\right)^{-2} C_{11.2}\right\} + \lambda^2(1-\alpha)^2\beta'_1 \right.$$
$$\left.\left[C_{11.2} + \lambda(1-\alpha)I_{p_1}\right]^{-2}\beta_1 + \lambda^2(1-\alpha)^2\beta'_2\left[C_{22.1} + \lambda(1-\alpha)\right]^{-2}\beta_2\right].$$

4.3 Naive Elastic net and Elastic net

4.3.1 Naive Elastic net (nEnet)

So far we have discussed naive elastic net. This estimator satisfies the properties raised earlier. Zou and Hastie (2005) say it is not satisfactory unless α is close to 0 or 1. That why they called it naive.

Note that ridge regression based prediction is accurate through bias-variance trade off. They found this conclusion due to the orthogonality assumption for their analysis. In the orthogonal case, we have

$$\tilde{\boldsymbol{\beta}}_n^{ridge}(\lambda) = (I_p + \lambda I_p)^{-1}\tilde{\boldsymbol{\beta}}_n, \tilde{\boldsymbol{\beta}}_n \text{ is LSE}$$
$$= \frac{\tilde{\boldsymbol{\beta}}_n}{1+\lambda}. \tag{4.6}$$

Alternatively, consider the naive elastic net case, the estimator is

$$\tilde{\boldsymbol{\beta}}_n^{ridge}(\lambda, \alpha) = (I_p + \lambda I_p)^{-1}\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha), \tag{4.7}$$

where

$$\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha) = (\frac{\sigma}{\sqrt{n}}\,\mathrm{sgn}(Z_j)(|Z_j| - \lambda\alpha)^+)|j = 1, \ldots, p)'$$
$$= \frac{\sigma}{\sqrt{n}}\,(\tilde{\beta}'_{1n}, 0')',$$

applying a family of diagonally orthogonal linear projections. Notice that (4.7) is nothing but a ridge regression estimator of a LASSO estimator which provides subset selection. Then, we can describe the procedure as: select the subset, then shrink it by ridge factor to obtain the nEnet. Zou and Hastie (2005), suggested Enet by rescaling nEnet by the inverse of ridge factor due to double shrinkage. Looking at (4.6) we find the scaled ridge as

$$(1+\lambda)\tilde{\boldsymbol{\beta}}_n^{ridge}(\lambda) = \tilde{\boldsymbol{\beta}}_n.$$

Consider (4.7) and scale it

$$(I_p + \lambda(1-\alpha)I)\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda) = \hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha).$$

Optimality property of naive elastic net destroyed.

4.3.2 Elastic net (Enet)

We have so far considered nEnet. Zou and Hastie (2005) point out it is not satisfactory except for the cases α near 0 or 1. That is why it is naive.

Note that ridge regression based prediction performs well through bias-covariance trade-off and it is a shrinkage estimator. Let us look at nEnet expression

$$\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha) = [I_p + \lambda\ (1-\alpha)C_n^{-1}]^{-1}\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha),$$

where

$$\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha) = (\sigma\sqrt{c^{jj}}\,\mathrm{sgn}(Z_{jn})\,(|Z_{jn}| - \lambda\alpha)^+\,|j = 1,\ldots,p)'.$$

Notice that

(i) $\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha)$ selects subsets of variables.
(ii) $[I_p + \lambda\ (1-\alpha)C_n^{-1}]^{-1}$ shrinks the variables of the subset selected, which means there is double shrinkage. Thus, Zou and Hastie defined elastic net (Enet) by rescaling $\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha)$ as given below:

$$\left[I_p + \lambda\ (1-\alpha)C_n^{-1}\right]\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha) = \hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha).$$

We define $\hat{\boldsymbol{\beta}}_n^{Enet}(\lambda,\alpha)$ by

$$\hat{\boldsymbol{\beta}}_n^{Enet}(\lambda,\alpha) = [1 + \lambda\ (1-\alpha)]\,\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha);$$

$$\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha) = \left(\beta_{1n}'\left[I_{p_1} + \lambda(1-\alpha)C_{11.2}^{-1}\right]^{-1}, \beta_2'\left[I_{p_2} + \lambda(1-\alpha)C_{22.1}^{-1}\right]^{-1}\right)'.$$

The bias and risk of $\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha)$ are given by

$$\mathrm{Bias}\left(\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha)\right) = \left(-\lambda(1-\alpha)\beta_1'\left[I_{p_1} + \lambda(1-\alpha)C_{11.2}^{-1}\right]^{-1}, -\lambda(1-\alpha)\right.$$
$$\left.\beta_2'\left[I_{p_2} + \lambda(1-\alpha)C_{22.1}^{-1}\right]^{-1}\right)',$$

$$R\left(\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha);\boldsymbol{\beta}\right) = \sigma^2\left[\mathrm{tr}\left\{(C_{11.2} + \lambda(1-\alpha)I_{p_1})^{-2}\,C_{11.2}\right\} + \lambda^2(1-\alpha)^2\beta_1'\right.$$
$$\left.[C_{11.2} + \lambda(1-\alpha)I_{p_1}]^{-2}\,\beta_1 + \lambda^2(1-\alpha)^2\beta_2'\left[C_{22.1} + \lambda(1-\alpha)\right]^{-2}\beta_2\right].$$

respectively. Thus, we obtain the risk of $\hat{\boldsymbol{\beta}}_n^{Enet}(\lambda,\alpha)$ as

$$R\left(\hat{\boldsymbol{\beta}}_n^{Enet}(\lambda,\alpha);\boldsymbol{\beta}\right) = \sigma^2\left[1 + \lambda(1-\alpha)\right]^2 R\left(\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha);\boldsymbol{\beta}\right).$$

This estimate does well compared to LASSO and ridge estimators as discussed by Zou and Hastie (2005).

Chapter 5

CONCLUSION

This thesis contains the study of three basic penalty estimators, namely, ridge regression, LASSO and elastic net based on the landmark contributions by Hoerl and Kennard (1970), Tibshirani (1996), and Zou and Hastie (2005) respectively. In their paper, mathematical analysis is based on Stein-type estimators. Later, Saleh and Sen (1978 - 1986) introduced preliminary test estimator in the class of shrinkage estimators using rank-based statistics. Saleh (2006) defines them as quasi-empirical Bayes statistics and introduced one-dimensional Stein-type estimator. To describe it, let β_{jn} be the jth regression parameter and $\tilde{\beta}_{jn}$ is the least-squares estimator. Then, we obtain the following shrinkage estimators

1) $\hat{\beta}_{jn}^{PT} = \tilde{\beta}_{jn}\ I\left(|\tilde{\beta}_{jn}| > \lambda\sigma\sqrt{c^{jj}}\right)$,

2) $\hat{\beta}_{jn}^{S} = \tilde{\beta}_{jn}\ \left(1 - \frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right) = \tilde{\beta}_{jn} - \lambda\sigma\sqrt{c^{jj}}\,\mathrm{sgn}\left(\tilde{\beta}_{jn}\right)$,

3) $\hat{\beta}_{jn}^{S+} = \tilde{\beta}_{jn}\ \left(1 - \frac{\lambda\sigma\sqrt{c^{jj}}}{|\tilde{\beta}_{jn}|}\right) I\left(|\tilde{\beta}_{jn}| > \lambda\sigma\sqrt{c^{jj}}\right)$.

Note that PTE is a discrete process giving 0 or $\tilde{\beta}_{jn}$ nothing in between. $\hat{\beta}_{jn}^{S}$ changes sign like James-Stein estimator. $\hat{\beta}_{jn}^{S+}$ results in correcting $\hat{\beta}_{jn}^{S}$ and uniformly superior to $\hat{\beta}_{jn}^{S+}$. The formula $\hat{\beta}_{jn}^{S+} = \mathrm{sgn}(\tilde{\beta}_{jn})\ \left(|\tilde{\beta}_{jn}| - \lambda\sigma\sqrt{c^{jj}}\right)^{+}$ is the traditional form of LASSO in the literature proposed by Tibshirani (1996). We defined $\hat{\beta}_{jn}^{S+}$ as positive-rule Saleh-type estimator and $\tilde{\beta}_{jn}^{S}$ as Saleh-type estimator.

Further, we find the risk expressions as

1) $R\left(\hat{\beta}_{jn}^{PT};\beta_j\right) = \sigma^2 c^{jj}\left[1 - H_3\left(\lambda^2,\Delta_j^2\right) + \Delta_j^2\left(2H_3\left(\lambda^2,\Delta_j^2\right) - H_5\left(\lambda^2,\Delta_j^2\right)\right)\right]$,

2) $R\left(\hat{\beta}_{jn}^{S};\beta_j\right) = \sigma^2 c^{jj}\left[1 - \frac{2}{\pi}(2\ \exp^{-\frac{\Delta_j^2}{2}} - 1)\right]\ ,\ \lambda_{opt} = \sqrt{\frac{2}{\pi}}$,

3) $R\left(\hat{\beta}_{jn}^{S+};\beta_j\right) = \sigma^2 c^{jj}\left[1 - H_3\left(\lambda^2,\Delta_j^2\right) + \Delta_j^2\left(2\ H_3\left(\lambda^2,\Delta_j^2\right) - H_5\left(\lambda^2,\Delta_j^2\right)\right)\right]$
$+\lambda^2\left(1 - H_1\left(\lambda^2,\Delta_j^2\right)\right) - 2\ \lambda\ \left(\phi(\lambda - \Delta_j) + \phi(\lambda + \Delta_j)\right)$
$= R\left(\hat{\beta}_{jn}^{PT};\beta_j\right) + \left[\lambda^2\left(1 - H_1\left(\lambda^2,\Delta_j^2\right)\right) - 2\ \lambda\ \left(\phi(\lambda - \Delta_j) + \phi(\lambda + \Delta_j)\right)\right]$,

where $H_\gamma(\lambda^2;\Delta_j^2)$ is the cdf of a non-central chi-square distribution with γ D.F. and non-centrality parameter $\frac{\Delta_j^2}{2}$. Similar result is available using cdf / pdf of N(0, 1).

Next, in our study, we found that the naive elastic net for regression model is given by

$$\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha) = \left[I_p + \lambda(1-\alpha)C^{-1}\right]^{-1}\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha),\ 0 < \alpha < 1,$$

where

$$\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha) = \left(\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha)|j = 1,\ldots,p\right) \text{ and } \left[I_p + \lambda(1-\alpha)C^{-1}\right]^{-1}$$

is the ridge factor in ridge regression.

We now state the findings of Tibshirani (1996) as follows:
(a) Ridge regression is a shrinkage estimator only, it does not select subset. If there are large number of small effects, ridge regression does best by a good margin, follows by LASSO and then by Preliminary test selector.
(b) LASSO does estimate and select subset simultaneously. It does best with LASSO when small to moderate number of moderate-size effects, followed by ridge regression, the preliminary test selectors.
(c) Preliminary test selector does best for small number of large effects, LASSO not so well and ridge regression quite poorly. As observed by Zou and Hasti (2005).
(d) For usual $n > p$, if there are high correlations between predictors, as by Tibshirani (1996), LASSO is dominated by ridge regression.

(e) In case of $p > n$ (high - dimension), LASSO selects at most n variables, LASSO is not well defined unless the bound on L_1-norm of the coefficients is smaller than a certain value.
(f) If there is a group of variables among which the pairwise correlations are very high. LASSO tends to select one variable from each group without caring which one it is.

In the thesis, we tried to put together all aspects of ridge, LASSO, and elastic net - penalty estimators in relation to the comments (a)-(e).

We expressed the naive elastic net solutions

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^{nEnet}(\lambda,\alpha) &= [1+\lambda(1-\alpha)C^{-1}]^{-1}\hat{\boldsymbol{\beta}}_n^{LASSO}(\lambda\alpha), \quad 0<\alpha<1 \\ &= \begin{pmatrix} [I_{p_1}+\lambda(1-\alpha)C_{11.2}^{-1}]^{-1}\tilde{\boldsymbol{\beta}}_{1n} \\ 0 \end{pmatrix},\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{1n}$ is p_1-vector of LSE and $[I_{p_1}+\lambda(1-\alpha)C_{11.2}^{-1}]\tilde{\boldsymbol{\beta}}_{1n}$ is the ridge regression estimator which is an improved estimator than $\tilde{\boldsymbol{\beta}}_{1n}$ and nEnet improves over Enet using efficiency criterion.

We obtain ridge regression using partial penalty function $k\boldsymbol{\beta}_2'\boldsymbol{\beta}_2$ which may useful for high dimensional estimation ($p > n$).

References

Breiman, L., (1996). Heuristics of instability and stabilization in model selection, The Annals of Statistics, 24, 23502383.

Bashtian, M. Hassanzadeh, M Arashi, and S. M. M Tabatabaey.(2011). Ridge Estimation Under the Stochastic Restriction. Communications in Statistics - Theory and Methods 40.21 37113747.

Donoho, David L and Johnstone, Jain M, (1994). Ideal spatial adaption by wavelet shrinkage, Biometrika, vol. 81, no. 3, pp.425 455.

Efron, B., Morris, C. (1973). Combining Possibly Related Estimation Problems. Journal of the Royal Statistical Society. Series B (Methodological), 35(3), 379-421.

Gao, X., Ahmed, S.E., and Feng, Y., (2017). Post selection shrinkage estimation for high dimensional data analysis. Applied Stochastic Models in Business and Industry, 33, 97-120.

Golub, G. H., Heath, M., and Wahba, G., (1979). Generalized Cross-Validation as a method for choosing a good ridge parameter, Technometrics, 21(2), 215-223.

Hoerl, A.E., (1962) Application of ridge analysis to regression problems, Chemical Engineering Progress, 58(3), 54-59.

Hoerl, A.E. and Kennard, R.W., (1970) Ridge regression: biased estimation for nonorthogonal problems, Technometrics, 12, 5567.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F., (1975). Ridge Regression: some simulations, Communications in Statistics-Theory and Methods, 4, 105123.

James, W., and Stein, C., (1961). Estimation with quadratic loss. Proceeding Berkeley Symposium on Mathematical Statistics and Probability, University of California, 1, 361379.

Kibria, B. M. G. and Saleh, A. K. Md. E, (2003). Preliminary test ridge regression estimators with Students t errors and conflicting test-statistics, Metrika, vol. 59, no. 2, pp.105-124.

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. Communications in Statistics-Simulation and Computation, 32, 419-435.

Kibria, B. M. G. and S. Banik (2016). Some Ridge Regression Estimators and Their Performances. Journal of Modern Applied Statistical Methods. 15 (1), 206-238.

Kibria, B. M. G. and Lukman, A. F. (2020). A new ridge-type estimator for the linear regression model: Simulations and Applications. Scientifica, Article ID 9758378, 1-16.

Lawless, J. F. and Wang, P., (1976). A Simulation Study of Ridge and Other Regression Estimators, Communications in Statistics - Theory and Methods, 5:4, 307-323.

Saleh, A. K. Md. Ehsanes; and Sen, Pranab Kumar. (1978).Nonparametric Estimation of Location Parameter After a Preliminary Test on Regression. The Annals of Statistics, 6(1), 154-168.

Saleh, A., (2006). Theory of Preliminary Test and Stein-type Estimation with Applications, John Wiley, New York.

Saleh, A.K.M.E., Arashi, M., and Tabatabaey, S.M.M., (2014). Statistical Inference for Models with Multivariate t-Distributed Errors, John Wiley, New Jersey.

Saleh, A. K. Md. E, Arashi, M. and Kibria, B. M. G. (2019). Theory of Ridge Regression Estimation with Applications, Wiley 2019.

Saleh, A. K. Md. E, Arashi, M., Saleh, R. and Norouzirad, M. (2021). Rank-Based Methods for Shrinkage and Selection with Application to Machine Learning, Wiley 2021.

Sen, P.K. and Saleh, A.K.M.E., (1979). Non parametric estimation of location parameter after a preliminary test on regression in multivariate case, Journal of Multivariate Analysis,9,322-331.

Sen, P.K. and Saleh, A.K.M.E., (1985). On some shrinkage estimators of multivariate location, Annals of Statistics, 13, 172 281.

Sen, P.K. and Saleh, A.K.M.E., (1987). On preliminary test and shrinkage m-estimation in linear models. Annals of Statistics, 15, 1580 1592.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58(1), 267-288.

Zou, H. and Hastie, T., (2005). Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society. Series B (Methodological), 67, 301320.

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