Assessing the Performance of Robust Penalized Regression Methods

By

Abdul Wahid

A DISSERTATION SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF PHILOSOPHY
IN STATISTICS

Supervised By
Dr. Zahid Asghar

Department of Statistics
Quaid-i-Azam University
Islamabad, Pakistan
2016.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Background of the Study

This thesis is concerned about multiple regression models, when a large number of explanatory variables are consider to explain the relationship between a response Y and a set of predictors $X_1, ..., X_P$. In this thesis, we wish to consider other suitable procedures rather than common least squares. We shall be detecting in further thesis, these alternative model building methods can supply preferable prediction accuracy and model that is easy understandable.

1.1.1 Prediction Accuracy

It is well known that the ordinary least squares often perform poorly in prediction on future data. When $n > P$, or we can say that the sample size(n) is greater than the total number of independent variables(P), in this situation the OLS estimates have small bias but high dispersion and provide more flexible model. However, if $P \gg n$, the numbers of explanatory variables are much larger than the numbers of observations, then OLS is highly variable, even infeasible, resulting over fitting and therefore poor predictions on test observations. In $P \gg n$ situation, then OLS coefficients estimates have no longer unique, the variance become boundless and leads to the OLS method cannot be utilized at all.

1.1.2 Model Interpretability

In more flexible and high dimensional, when $P \gg n$, regression models some or most predictors are actually not associated with the dependent variable. As well as such unrelated predictors advance to needless complication in the constructing model. We focus in this thesis to uses some methods by removing these fake variables and a model that is more easily interpreted. To deal with the challenges stated before, the constraint regression approaches, named shrinkage and regularization procedure too, have been established. However, diminishing several predictor coefficients towards 0, to do such can be produce bias estimates, but estimated coefficients will have little dispersion or standard deviation. This may consequences to raise prediction accuracy because the resulting model gives the minimum mean squared error. Model coefficients are condensing by attach a penalty on their magnitude, it is done to add a penalty function to usual residual sum of squares. Furthermore, several of these approaches e.g. the Least Absolute Shrinkage Selection Operator(R.Tibshirani,1996) enable variable discarding or choosing just as only the influential explanatory variables remain in the model(Szymczak, et al.2009).

1.2 Statement of the Problem

When P is much bigger than n, frequently inscribed $P \gg n$, such issues have happened to expanding importance, particularly in genomics and different fields of computational science and marketing. Previously we have known, the inflated dispersion and over fitting becomes vital issues in this framework. Some penalization techniques have been suggested to better OLS, for instance, ridge shrinkage method (Hoerl and Kennard,1988) reduces the residual sum of squares under to the $L_2 - norm$ of the slopes coefficients. It is a method to continuously reduce the slope coefficients, and gain its good prediction achievement between a bias-variance trade-off. However, ridge regression does not construct a parsimonious (or we can also say simple) model, and at all times retain total explanatory variables in the model. A better procedure called the, LASSO, was invented by Tabishrani(1996). The LASSO is a penalized least square method imposing an L_1-penalty on the regression coefficients. However, in $P \gg n$ correlated collectively predictors scenario, the LASSO is not the complete procedure(Efron et al.,2004). Alternative methods have been introduced to deal with this problem. The motive of this work is to assess the numerical performances of ridge regression, LASSO, Elastic Net and proposed method.

1.3 Research Motivation

The motivation for using penalized regression is that in the multiple regression and high dimensional data problem, the ordinary least squares method are not appropriate. The variable selection is important and fundamental to high-dimensional regression models for increasing predictability and select significant variables. Many approaches are proposed such as, stepwise predictor choosing methods, these approaches are computationally very laborious and infeasible when P, numbers of predictor variables, become sufficiently large, and another drawback is that these methods ignore stochastic error terms. In this thesis penalized regression methods for regularization and variable selection are compare, such like ridge regression, lasso, elastic net, adaptive lasso and proposed by us, adaptive lasso with ridge coefficients are used as weights.

1.4 Aims and Objectives of the Study

The main aim of this research is to assess the performance and advantages of using Ridge, LASSO, Elastic Net, Adaptive Lasso and proposed Adaptive Lasso methods in multiple regression models in different situations. We hope to achieve this aim through the following objectives:

I. Application of penalized regression methods in robust regression.

II. Identifying the methods that have better prediction performance.

III. Identifying the methods that have oracle properties, in predictor variable selection.

IV. The performance of these regularization methods in generalized linear models.

V. Assessing the functions of stated procedures, whenever the error expression has non constant variance.

1.5 Significance of the Study

Why the penalized regression methods are preferred over classical least squares and subset selection techniques in multiple linear regression models. The first reason is that the least squares estimates often have low bias but high variance, even not unique in $P \gg n$ case. Prediction accuracy can occasionally be enhanced by condense or setting various co-efficients to 0 by using penalization regression methods. By working as such we immolated a negligible amount of bias to decrease the dispersion of the estimated values, subsequently may increase the total prediction accuracy, it is called Bias-Variance trade off. Secondly, With a large number of predictors, we often would like to determine a smaller subset that exhibit the strongest effects. In order to get the big picture, we are willing to sacrifice some of the small details. In subset selection, by retaining a sub set of the predictors and discarding the rest, subset selection produces a model that is interpretable and has possibly lower prediction error than the full model, or the model obtained by using least squares approach. But, these are discrete processespredictors are either held or discardedit usually shows high fluctuations, and results to increase the test error of the full model. Penalized regression approaches are more continuous, and does not hurt as much from high variability. We also show why penalized methods are preferred, over each other when we faced the different scenario. In achieving this, we discussed and compared different penalized methods used in different situations. This work is also aimed at providing assistance to researchers to ease their decision making as to which technique to be used when encountered with the problem of different scenarios.

1.6 Scope and Limitations of the Study

Regularized and variable selection regression techniques for linear models have been developed in final two decagons to solve the shortcomings of ordinary least squares and subsets regression procedures with regard to prediction accuracy and visualization. Particularly, whenever we face the problem of high dimensionality. These approaches are widely implemented in the field of computational biology, genes study, finance and marketing, etc. One essential and well known drawback of regularization approaches is to producing biased estimators of model parameters.

These methods have also complicated computations in proving various large sampling and oracle properties.

1.7 Robust Regression

Ordinary least squares (OLS) estimates work not better whenever the distribution of ϵ_i factor is not guassian, specially when the error is symmetric or asymmetric but heavy-tailed. Additionally, estimator performance is considered under conditions in which errors follows different non-normal distributions. One treatment is to discard extreme observations and apply ordinary least squares, such as trimmed least squares, but mostly we cannot ignore these influential observations because these are parts of real data set. Another most widely applied field, termed robust regression, is to employ a fitting criterion that is not as vulnerable as least squares to unusual data. The well known general technique of robust regression is M-estimation, introduced by Huber (1964). The study addresses the problems of penalized regression approaches in the situation of outliers on the multiple linear regression models.

1.8 Generalized Linear Models

Often in linear regression models it is suppose that the response Y is quantitative. However in several cases, the dependent variable Y is qualitative. For instance, marital status is qualitative variable, taking on values married or unmarried. In early, logistic regression were used in the epidemiological research and nowadays it is commonly employed in fields ranging from engineering to finance and marketing, are settled in the class of probability models. Generally qualitative variables are named as categorical, we will utilize the mentioned terms interchangeably. In logistic regression we learn techniques for estimating categorical dependent variables. Analogous to penalized regression, where the objectives was to minimize a penalized residual sum of squares, we may incorporate a constraint term to the log-likelihood function for the same purpose of achieving a well built, and furthermore correct regression model for high-dimensional data. In this thesis we also study the penalized logistic regression approach for Poisson logistic regression.

1.9 Organization of Thesis

The thesis is divided into six chapters. Following this introductory chapter is *chapter*2, which deals with some related litrature, methodology and diagnosis. *Chapter*3 presents the new method about variable selection and visualization. The first part is discusses the properties and some characteristics of the propose method. The second part covers the simulation study, and third part real data performance. *Chapter*4 elucidates the role of these regularization and model selection procedures in case of non-constant error term variances, called heteroscedasticity. *Chapter*5 deals with the performance of proposed method versus other regularization methods in generalized linear models. A brief conclusion is drawn in *chapter*6. Following this concluding chapter is *appendixA*.

Chapter 2

Literature, Methodology and Regression Diagnostics

2.1 Introduction

In this chapter we present some related literatures on the works of different scholars regarding the penalized regression and their properties. Many authors have proposed several penalized regression methods to beat the defects of ordinary least squares and subset selection of variables methods with regards to prediction accuracy and variables selection. The extensive literature on selecting subsets of predictor variables was well reviewed by Hocking (1976). Thompson (1978) has also reviewed subset selection in regression. A.J .Miller (1984) was also discussed the selecting of subsets of regression variables and computational algorithm. Subsets selection methods construct a model which is interpretable and has mostly minimum test error than the complete model, but these approaches are discrete in predictor determine, mostly presents high dispersion, leads to increase in test error of true model. Penalized shrinkage methods are more continuous, and do not suffer as much from high variability.

2.2 Penalized Regression

It is a common fact that, OLS frequently does poorly in both prediction and clarification especially when number of predictor variables(P) become greater than the numbers of observations(n). Hoerl, and Kennard (1970) introduced the ridge regression which estimates the regression coefficients through an $L_2 - norm$ penalized least-squares criterion. As a uninterrupted shrinkage method, ridge regression attains good prediction presentation across a bias-variance trade-off. Though, ridge regression don't construct a simple model which will be easy in explanation. Best subset selection in contrast produces a sparse model, but it is extremely variable because of its inherent discreteness, as addressed by Breiman (1996). Frank and Friedman (1993) initiated bridge regression that curtails the Residual Sum of Squares (RSS). The estimator from bridge regression is not explicit. A favorable method named the LASSO was suggested by Tibshirani (1996). LASSO is a penalized residual sum of squares technique imposing an $L_1 - penalty$ on the predictor coefficients. Due to the type of $L_1 - norm$, the LASSO does both continuous shrinkage and self-executing predictor selection simultaneously. As we know that the LASSO has proved advantages in various scenarios, but it has some drawbacks, describe below. (a) In $P \gg n$ situation, the LASSO pick no more than n(i.e, sample size) variables before permeates, due to the kind of the convex optimization problem. This appears to be restrictive characteristic for a predictor picking method. Additionally, the LASSO is not unambiguous unless the bound on the $L_1 - norm$ of the coefficients is smaller than a definite value. (b) Whenever there is a class of variables among which the pairwise relationships are much high, next LASSO moves to sets only one predictor from the group and doesn't control which one is preferred. (c) In familiar, $P < n$ condition, if there is a situation of high relationships among features, it has been numerically found that the test error of the LASSO is greater than ridge regression (Tibshirani, 1996). Situations (a) to (c) form the LASSO an unsuitable feature selection technique. Fan and Li (2001) suggested the Smoothly Clipped Absolute Deviation (SCAD) constraint for predictor selection to minimize bias and support some conditions to produce continuous solutions. Also, they derived the constant shrinkage parameter and large sample distribution of the proposed estimator and provide that it follows the oracle property (predictors should be consistently choosing). Large sample properties of SCAD estimators are studied in Kwon and Yongdai (2012). But, this formula is non-convex, it is a limitation because it leads to computational issues and make solution difficult. Efron et al., (2004) later suggested Least Angle Regression Selection (LARS) for a model selection algorithm. They proved that with a simple modification, the LARS algorithm implements the LASSO. Efron et al., (2004) also studied an efficient way of choosing the best fit and the effective degrees of freedom of the LASSO, where it was initiated that, the size of the active set (the pairs of quantity corresponding to explanatory variables to be selected) can be utilize as a measure of the degrees of freedom, that vary, may not be monotonically, along with coefficients paths of LARS. Zou et al.,(2007) developed the work of (Efron et al.,2004) and produced that the number of nonzero estimates is an unbiased estimate for degrees of freedom of the LASSO. Furthermore, Zou et al.,(2007) proved that the unbiased estimator is asymptotically consistent, therefore many model selection approaches can be utilized with the LARS formula for the best LASSO fit. Zou and Hastie (2005) proposed the Elastic Net penalty which is based on combined penalties of LASSO and ridge regression. The penalty parameter determines how much weight should be given to either the LASSO or ridge regression. The Elastic Net with set to 0 is equivalent to ridge regression. The Elastic Net with close to 1 performs much like the LASSO, but removes any degeneracies and odd behavior caused by high correlations. Zou and Hastie (2005) highlighted two aspects that are important when evaluating the quality of a model: (a) Prediction results on future data; it is hard to clarify a model that predicts better; (b) Explanation of the model; scientists favour a model that have minimum predictors because it place more little on the relationship between the response and covariates. Parsimony is particularly a major matter if the number of explanatory variables are outsize.The elastic net produces a sparse model with good prediction accuracy, while encouraging a grouping effect. Zou and Hastie shows that elastic net is superior method over LASSO and other shrinkage methods. The Dantzig selector which was introduced by (Candes and Tao, 2007), a slightly modified version of the LASSO. Both were named the solution, Dantzig selector (DS). George Dantzig also demonstrate a number of fascinating mathematical properties for his technique, but the similar properties were proved for the LASSO too, as set out after by Bickel et al. (2008).

Thus, this method gives same results as LASSO. Unfortunately, the mathematical properties of DS technique are somewhat not satisfactory . The grouped LASSO procedure was suggested by Bakin(1999) and later developed by Lin and Zhang (2006), measured and generalized by Yuan and Lin (2007). If the predictors belong to some pre-defined groups, then grouped LASSO may be shrunk and select the variables of a group together. Hui Zou (2005) preferred a fresh kind of the LASSO, and give named *adaptive Lasso* in his method he used the OLS estimated coefficients as weights in the L-1 penalty expression. He also prove that the this adapted method follow the oracle properties. In high dimensional problems then it is nontrivial to find the weights for adaptive Lasso. Then Jian Huang et al. (2008) discussed the adaptive Lasso for high dimensional regression problems under restrictive a partial orthogonality condition which often not met in high dimensional problems. Copas (1983), Frank and Friedman (1993) have been done the comparisons of different penalized and shrinkage procedures.

2.3 Regularization and Shrinkage Methods

2.3.1 Introduction

Classical linear regression assumes that the dependent variable vector $y = (y_1, ..., y_n)^T$ is a straight line combination of P predictors $X_1, ..., X_P$ and an unknown parameter vector $\beta = (\beta_1, ..., \beta_P)^T$ and also an extra error expression $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$. Thus, ordinary regression model are as under

$$y = X^T \beta + \varepsilon \tag{2.1}$$

where ε has normally distributed with 0 mean and constant variance σ^2 and the outline matrix $X = (X_1, ..., X_P)$ which is established on n independently and identically distributed samples. The dependent variable y is centered and the explanatory variables are standardized. Thus:

$$\frac{1}{n}\sum_{i=1}^{n} y_i = 0, \frac{1}{n}\sum_{i=1}^{n} X_i = 0; \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2 = 1, \tag{2.2}$$

for j=1,2,...,p Since the response is centered and the predictors are standardized, no intercept has to be estimated. The usual estimation procedure for the parameter vector β is the minimization of the residual sum of squares with respect to β:

$$\tilde{\beta}_{OLS} = argmin_{\beta}(y - X\beta)^T(y - X\beta) \tag{2.3}$$

Then, the ordinary least squares (OLS) estimator is $\tilde{\beta}_{OLS} = (X^TX)^{-1}X^Ty$. A disadvantage of this ordinary estimation technique has main deficiency of feature selection. Even slope estimators whose corresponding independent features have leaving or low impact on the dependent variable y stay in the regression model. If we have a regression model with a lot of explanatory variables, then we want to determine a parsimonious model which is simple to elucidate and low variance. Actually, the subset selection procedures builds sparse models but they are highly variable techniques due to discreteness. To solve these difficulties, construct regression models via regularization and penalization techniques were suggested. The regularization techniques are depend on penalty expressions and should supply unique estimates of the unknown vector β. Moreover, better results of the prediction accuracy can be obtained by shrinking the coefficients or setting some of them to zero. So that we get regression models which should include only the strongly related factors and that are easy to understand. In the following, An outline of various already built visualization and shrinkage methodologies and modified technique are as under,

2.4 Penalized Least Squares Approaches

Shrinkage and feature selection approaches for regression problems are established on penalized least squares

$$L(\lambda_n, \beta) = \{(y - X\beta)^T(y - X\beta)\} + P(\lambda_n, \beta) \tag{2.4}$$

and the estimates of the unknown vector β are derived by minimizing the following equation,

$$\tilde{\beta} = argmin_{\beta}\{L(\lambda_n, \beta)\} \tag{2.5}$$

The penalty expression $P(\lambda_n, \beta)$ rely on the tuning parameter λ_n which governed the step-down intensity. For the regularization parameter $\lambda_n = 0$ then we get the simple least squares solution. On the opposite, for high values of λ_n the impact of the penalty expression on the coefficient estimates grows. Thus, the penalty part regulates the belongings of the estimated unknown constant vector β, while attractive characteristics are predictor selections.

2.4.1 Ridge Regression

Ridge Regression is clearly different from ordinary multiple linear regression whose goal is to circumvent the issue of instability arising, amongst others, from collinearity of the predictors variables. As PCA and PLS build uncorrelated linear combinations of the explanatory variables according to handle a problem of multicollinearity. This method, on contrary, works with the original variables and tries to minimize a penalized residual sum of squares [14, p.63]

$$\tilde{\beta}^{ridge} = argmin_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\} \tag{2.6}$$

where $\lambda \geq 0$ is the complexity parameter, also referred to as regularization parameter, that control the amount of shrinkage. Like OLS, ridge regression all the predictor variables, but typically with smaller coefficients, depending upon the value of tuning parameter. As can be seen that (5), $\lambda = 0$ corresponds to the OLS regression. To further illustrate this, the equation may be equivalently written as objective function which minimizes least squares sum and constraint which is the restricted sum of the squared coefficients. In practice, no penalty is applied to the intercept β_0, and variables are scaled to ensure invariance of the penalty term to the scale of the original data. It can be shown, assuming centered data and estimating β_0 by the mean of response variable y, that the solution to equation (5), in the following written in matrix framework,

$$RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \tag{2.7}$$

is given by,

$$\tilde{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{2.8}$$

The capital I display the identity matrix. Adding λ to the diagonal of $X^T X$ makes the problem non-singular, even the multicollinearity is present. Further insight into the nature of ridge regressions is gained by employing singular value decomposition. Let $X = ULV^T$ be SVD of the centered $n \times P$ design matrix X, where U and V are the $n \times P$ and $P \times P$ orthonormal matrices and L is a matrix whose nonzero elements only on diagonal and of order $P \times P$ consisting of the singular values of X ordered by value. So, we may rewrite the least squares equation

$$X\tilde{\beta}^{ls} = X(X^T X)^{-1} X^T y \tag{2.9}$$

to

$$X\tilde{\beta}^{ls} = UU^T y \tag{2.10}$$

The ridge gives result as following,

$$\tilde{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{2.11}$$

$$= UL(L^2 + \lambda I)^{-1} LU^T y \tag{2.12}$$

$$= \sum_{j=1}^{p} u_j \frac{l_j^2}{l_j^2 + \lambda} u_j^T y \tag{2.13}$$

where the u_j are the column of U and l_j^2 are the diagonal entries of L^2. According to above equation, it reduces the coordinates of response vector with respect to the orthonormal basis of matrix U by the factor $\frac{l_j^2}{l_j^2+\lambda}$, which implies that more quantity of contraction is concerned to basis vectors which compact eigenvalues l_j^2.

2.4.2 LASSO Regression

The Least Absolute Shrinkage and Selector Operator (LASSO) introduced by Tibshirani(1996). This technique reduces the coefficients of many of the variables not simply towards 0 like ridge regression, but exactly to 0, giving an implicit form of variable selection. More formally, the method minimizes residual sum of squares subject to constraint on the sum of absolute values (L_1 penalty term) of the regression coefficients $\sum_{j=1}^{p} |\beta_j| \leq t$, which can be equivalently written as [14, p.68],

$$\tilde{\beta}^{lasso} = argmin_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\} \tag{2.14}$$

Here L_1 penalty term is used in place of L_2 penalty in equation (5).

As per definition of LASSO, forming the constant t adequately small in the constraint leads to several coefficients set to be completely zero. On the other hand if t is chosen large enough, the LASSO estimates equal the least squares solution, thus LASSO does a kind of continuous subset selection. Further insight into the shrinkage behavior can be gleaned from the orthonormal design case. Let X be a $n \times P$ design matrix and suppose that $X^T X = 1$. Then it can be easily shown that solution to the above equation have following form,

$$\tilde{\beta}_j^{lasso} = sign(\tilde{\beta}_j^{ls})\{|\tilde{\beta}_j^{ls}| - \gamma\}^+ \tag{2.15}$$

where γ is determined by the state, $\sum_{j=1}^{p} |\tilde{\beta}_j^{lasso}| \leq t$ and $\tilde{\beta}_j^{ls}$ is the OLS estimates.

In the first part of this subsection it was argued that the LASSO shrinks coefficients in the general (non-orthonormal) setting exactly to zero, in contrast to ridge regression. At least for two dimensional case, figure 2.1, will provide some visual insight, whether this statement holds. The criterion $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$equals quadratic function,

$$(\beta - \tilde{\beta}_j^{ls})X^T X(\beta - \tilde{\beta}_j^{ls}) \tag{2.16}$$

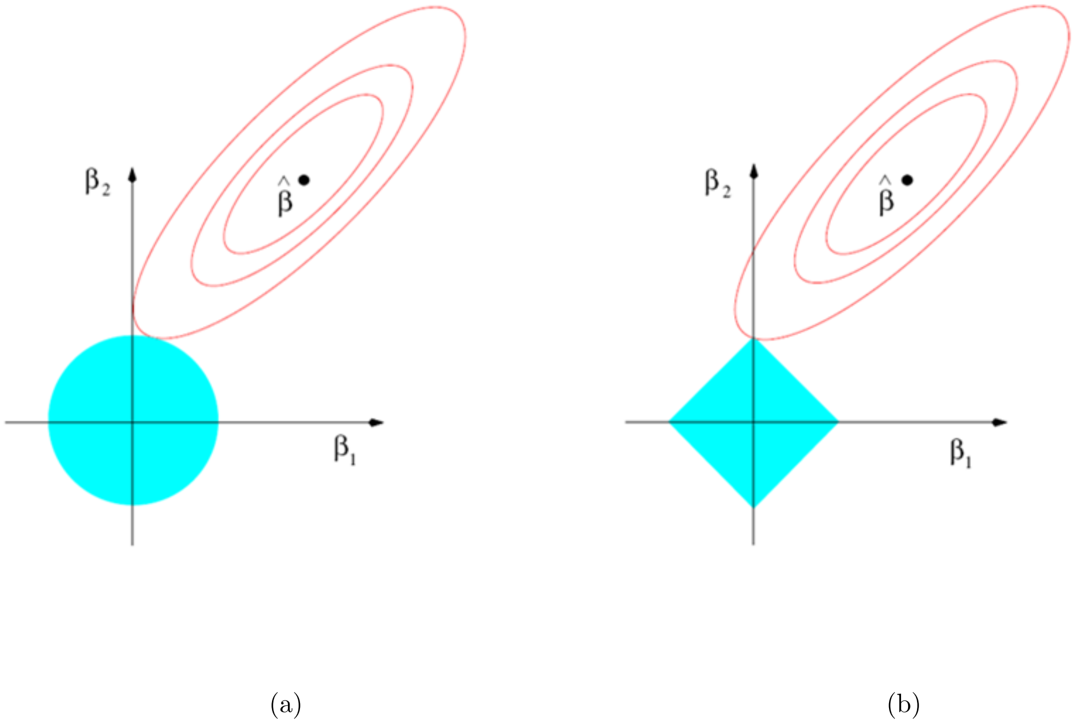


Figure 2.1: Estimation picture for the ridge regression (left) and LASSO (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.[14, p.71]

plus a constant error term, which may simply be derived from the normal equations. The elliptical contours of this function are shown by the full curves in figure 2.1, centered at the least squares estimates. The constraint regions are rotated square (L_1 Lasso penalty) in 2.1(b), and the circle (L_2 ridge penalty) in 2.1(a). The Lasso results is the first spot that the contours meet the square, and that will happen sometimes at a corner, corresponding to a zero coefficient, whereas in ridge regression there are no corners for the contours to hit, thus zero solution is not occur.

2.4.3 Elastic net Penalized Rgression

Latterly, Zou and Hastie(2005) have recommended a fresh constraint regression method, named elastic net, to inscription the poor functions of LASSO in many correlated explanatory variables condition. Just as LASSO, elastic net can assembles sparse models plus condense slope coefficients. Nevertheless, if there is a group of extremely correlated regressors, this method incorporates overall the X variables in the group with enhance prediction accuracy. In case of $P \gg n$, the LASSO don't work good and elastic net performing better as we have discussed.This procedure combined the LASSO and ridge penalties, mathematically written,

$$\tilde{\beta}_{ELN} = argmin_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j) + (1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2\} \tag{2.17}$$

where $0 \leq \alpha \leq 1$ is the penalty parameter. When $\alpha = 1$ then elastic net gets simply ridge regression and $\alpha = 0$ then it provide the convex LASSO penalty.

2.4.4 Adaptive LASSO

As we know that the LASSO solution does not fulfill the oracle properties. To overcome this problem, an adaptive Lasso was suggested by H.Zou, et.al(2006). He has present that mentioned method an oracle procedure. Because the LASSO puts the $\beta' s$ to be identically penalized in the L_1 penalty, and adaptive LASSO allocate OLS coefficients as weights to last expression. The adaptive LASSO penalized sum of squares defined as [42, p.1420],

$$\tilde{\beta}_{AL} = argmin_{\beta}||y - \sum_{j=1}^{p} x_{ij}\beta_j||^2 + \lambda\sum_{j=1}^{p} w_j|\beta_j| \tag{2.18}$$

Here w_j is a given weights vector. We display that if the weights are arranged and sensibly chosen, later the adaptive lasso can have the oracle properties.

2.4.5 Adaptive Penalized Least Squares

The adaptive LASSO uses the ordinary least squares estimates as weights. But it is infeasible to find adaptive LASSO estimates for high dimensional data problems($p > n$). Therefore I suggested to use the ridge regression coefficients as weights instead of OLS estimates. The proposed penalized criterion is same as adaptive LASSO except w_j.

2.5 Regression Diagnostics

In the previous section I introduced procedures for the purpose of developing good predictive relationships between response variable and predictors. The task of validating such predictive capabilities and issues addressing model selection and assessment, shall be discussed in the following. Model selection has to be understood as an estimation procedure for determining the achievement of various models in order to select; the perfect one. In the LASSO case that would simply run to a series of estimations over a grid of regularization parameter values λ, and the best λ according to some error criterion is chosen. The assessment of the resulting model is then performed on new data (out-of-sample testing) by calculating the prediction error. A common approach in data-rich situation is to randomly split the data set into three subsets: a training set for fitting, a validation set [1] to evaluate test error for model choosing and the holdout data part for estimation of the generalization error (prediction error over an independent subsample). Such scheme would bypass the circumstance that in case of using the same data for fitting and testing, the prediction error would be underestimated substantially, and usually a tendency towards over fitting the model can be observed (it will be discuss in next subsection). Coming up with a general, dividing rule is far from being an easy job, by reasons of, among other things, the dependence on the signal-to-noise ratio(i.e, SNR) in the data set, the training sample size and the complexity of the model. For instance, a typical split might be 50 percent of the data used in training part and 25% for validation and assessing respectively. Besides quantitative issues, like the question of how many observations each

[1] The validation step will be skkiped in the simulation part of this thesis, and I will just split the data set into a training set for model selection and a test set for model assesment.

subset should contain, one has to overcome qualitative problems concerning a proper representation of all classes in each subset. For instance, if we consider a binary response vector y of a logit model where each observation y_i, $i \epsilon I_n$ can be classified into responders $y_i = 1$ and non-responders $y_i = 0$, it will be essential to have a certain number of responders and non-responders in each subset, otherwise the model will not learn to classify correctly. Such strategies go by the name of stratification.

2.5.1 Training Error and Test Error

The proper assessment of a statistical learning technique on a specific data set, we demand a particular method to compute how well its predictions truly represents the observed data. In the regression setting, the most common measure is the mean squared error(MSE),or training error. Given the standard regression model $y = f(X) + \epsilon$ with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$ and a training set τ, the training error (loss over the training set) may be defined as [14, p.220],

$$\overline{err} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\hat{x_i})) \tag{2.19}$$

where $L(.,.)$ is some appropriate loss-function. This mean squared error will be minimum if the predicted response $f(\hat{x_i})$ are too near to the actual values of $f(x_i)$, and its value will be high for a part of the data set, if the predicted and actual of Y differ considerably. The error defined in (17) is determined utilize the training data set, this part of data is consider for fitting the regression model, and may be very accurately be introduce to as the training MSE. However, we are not interested in mean squared error calculated from training part of our data, or basically we do not want to assess our learning technique on already occur data, because something happened that is completely known. Therefore, we are focused in the test error rate which is obtained to implement the learning procedure to unobserved data set. Why is this what we think about. Assume that we received clinical measurements (e.g. weights, hypertension, exercise time, infertility, color blindness, food, etc) for sampled patients, and results about whether each patient has myocardial infarction. We can utilize these information of every patients and apply the statistical learning approach to predict risk of myocardial infarction consider on clinical measurements. In application, our main goal is to implement any learning method to precisely predict the stated disease risk for future unseen patients prior on their clinical measurements. Usually, anyone are not often wanted to whether or not the learning approach accurately predicts myocardial infarction risk for patients used to train the model, because we already know which of those patients have disease. The error finding on holdout set, also introduced to as generalization error, may be written as the prediction error over the test part,

$$error_{\tau} = E[L(y, f(\hat{x_i}))/\tau] \tag{2.20}$$

Note that the prediction error assigns to the error for certain training part of data, since τ is fixed. A associated quantity is the expected test error $Err = E[Err_{\tau}]$, averaging over everything that is random, including the training set. In case of a squared error-loss, the expected test error can be expressed with regards of bias and variance decomposition,

$$Err = E[y - f(\hat{x_i})] \tag{2.21}$$

or,

$$Err = \sigma^2 + [E(f(\hat{x_i})) - f(x)]^2 + E[f(\hat{x_i}) - E(f(\hat{x_i}))]^2 \tag{2.22}$$

or,

$$= \sigma^2 + Bias^2(f(\hat{x_i})) + Var(f(\hat{x_i})) \tag{2.23}$$

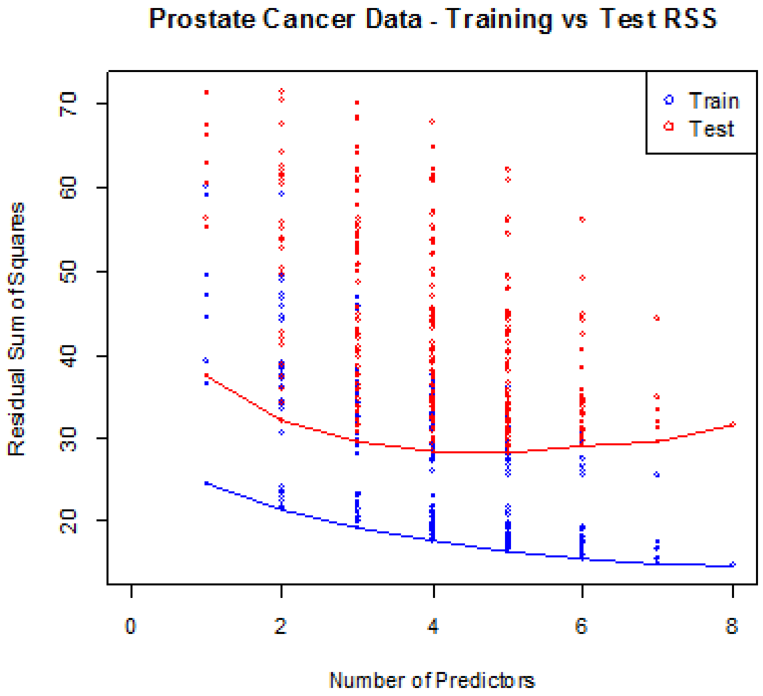


Figure 2.2: Comparison of test error and training error on prostate cancer data set.

The first term is the variance of error ϵ and cannot be avoided. The second term of above equation is equivalently to the squared of biasness, the quantity by that the sample mean of our estimate differ from the real mean, and the final expression represents the variation of our estimate. Typically, an increase in the complexity of a model results in a minimum bias but maximum variance, in the literature well known as the bias-variance trade-off. In other words, it is commonly said, that a model with large bias but low variance under fits the data, while a model with low bias large variance tends to over fits the model. However, for the sake of completeness, it should be mentioned that the behavior of the trade-off may depend on the chosen loss function. This is a elementary quality of learning method that support nevertheless of the special data set at hand and whatever of the statistical learning approach being applied. However, model flexibility(more predictors in the existing model) rises, training MSE will decline, in any case, the test MSE may not. This situation is called over fitting the data. This is shown in figure 2.2, we use the prostate cancer data set. Clearly, we can see that however the complication of any method grows, we get consistently decline in the MSE (blue curve), and the test MSE (red curve) is first decreases upto some extent, but model flexibility increase then test MSE also increases and make a U-shape curve. Various approaches have been established for the major goal of estimating the prediction error. Few of these techniques are introduced here, divided into theoretical methods [2] and empirical.

2.6 Theoretical Methods for Model Selection

Let us define with 'op' the optimum as the difference between the training error, $\overline{err}$ over the training data set $\tau = (y^{train}, X^{train})$ and the in-sample error, i.e Err_{in} over (y^{new}, X^{train}) [14, p.229],

$$op = Err_{in} - \overline{err} \tag{2.24}$$

The in-sample error rate is explained as [14],

$$Err_{in} = \frac{1}{n} \sum E_{y^{new}}[L(y_i^{new}, \hat{f}(x_i^{train}))/\tau] \tag{2.25}$$

[2] We donot consider theoretical methods in this thesis in the simulation part,but we discuss briefly in this section.

Its meaning that new response values y_i are observed at each of training points. Furthermore, the expected optimism ω is the prediction of the optimism over the training data set outcomes values with fixed predictors $\omega = E_y(op)$, and can generally be written as [14],

$$\omega = \frac{2}{n}\sum_{i=1}^{n} Cov(\hat{y_i}, y_i) \tag{2.26}$$

The last equation simply indicates that the stronger we fit the data, the larger the covariance between $\hat{y_i}$ and y_i will be, thus increasing the expected optimism ω. Upon taking the expectation of equation (20) we may rewrite as [14],

$$E_y[Err_{in}] = E_y[\overline{err}] + \frac{2}{n}\sum_{i=1}^{n} Cov(\hat{y_i}, y_i) \tag{2.27}$$

which in the linear case [3] for additive models $y = f(X) + \epsilon$ can be reduced to,

$$E_y[Err_{in}] = E_y[\overline{err}] + \frac{2d}{n}\sigma_\epsilon^2 \tag{2.28}$$

where d should capture the model complexity. The last expression is considered as the underlying idea of the information criteria introduced in the next subsection, which is first estimating the optimism and next adding it to the error calculated from the training observations to get an approximate for the in-sample error[4].

2.6.1 Akaike Information Criterion

The Akaike Information Criterion abbreviated as AIC was developed by Akaike in 1971 and proposed for the purpose of statistical identification as a measure of the goodness of fit of estimated models. Akaike suggested to take the log-likelihood estimate as a criterion of, fit, of a model, by reason of being a quantity which is most sensitive to deviations of the model parameters from the

[3] equation (23) hold for approximately in the non-linear case too.

[4] That in the in-sample error is not exactly brilliant is evident,since future values are not probable to match alongside their training set observations, but, especially for model testing, the in-sample mistakes is a appropriate alternative.

true values, and add a correction term(similar to the expected optimism in equations (23) and (24)).

$$AIC = -\frac{2}{n} ln(y_i, G(X\tilde{\beta})) + 2.\frac{d}{n} \tag{2.29}$$

If the maximum likelihood is identical for two competing models it will be the best choice, when focusing on the principle of parsimony, to take the less complex one, i.e., the model with fewer parameters. Since d in the last equation acts as a measure of the model complexity, or in Akaike words, d is the number of independent adjusted parameters, the model with the lowest AIC score should be preferred. Furthermore, AIC is by definition an asymptotically unbiased estimator of the mean expected log-likelihood, but is in general not asymptotically consistent in terms selection criterion, it is generate a class of models, consisting the true model, the chance that AIC will pick the correct model is strictly smaller than as the sample size tends to infinity. Typically, AIC tends to choose too complex models as the sample size increases, given room for further modifications, as for instance the Bayesian Information Criterion(BIC).

2.6.2 Bayesian Information Criterion

BIC was developed by Schwarz in 1978, and is applicable like AIC in situation where the attachment is done by optimization of a log-likelihood function.

$$BIC = -2ln(y_i, G(X\tilde{\beta})) + log(n).d \tag{2.30}$$

In spite of the similarity to AIC, (the component 2 is changed by log(n)), BIC is motivated differently. More precisely, it is motivated in a Bayesian framework, originating from approximating the evidence ratios of models known as the Bayesian factors. Another difference is that BIC enjoys the consistency property in terms of selecting the true model, but commonly preferred models that are manageable, due to substantial penalty on complicated terms. Another approach is C_p, defined as,

$$C_p = \frac{1}{n}(resd.SS + 2d_1\tilde{\sigma_1}^2) \tag{2.31}$$

In last equation $\tilde{\sigma_1}^2$ is an estimate of the error term dispersion. Like BIC, the C_p will movie to take on a little value for a ideal regression form with a minimum test inaccuracy, and normally we pick the model that has minimum C_p result.

2.7 Empirical Methods for Model Selection

Cross validation [4,14] and Bootstrapping [9, pp.1293−1296] are re-sampling methods for estimating the expected prediction error. Although being numerically intensive, and thus making re-sampling only applicable in an environment with access to fast hardware, its advantage lies in the conceptually simplicity, that is, re-sampling requires fewer assumptions and has a greater generalizability than traditional parametric approaches. We have done a lot of use of these two approaches in this thesis.

2.7.1 Cross Validation

In k-fold cross validation the observations set, further denoted by D, for simplicity, is randomly split into k mutually exclusive subsets, $D_1, D_2, ..., D_k$ of roughly in similar size. Then the model fitted k times. Each time $t = 1, 2, ..., k$ it is fitted on $\frac{D}{D_t}$ and for testing the prediction error is calculated when predicting D_t. The type of partitioning (the number of folds) allows a specific classification of cross validation. For instance, $k = n$ goes by the name of $leave - one - out$ cross validation(LOOCV), and here, as name the suggests, in each fit one observation is left out from the training data set. However, LOOCV indicates not only high computational costs, since n fits are necessary, but high variance in the cross validation estimate, because the D_i's for $i = 1, 2, ..., k$ are so similar to one another. It is interesting to note that the LOOCV technique is asymptotically equivalent to AIC, thus the later can be used as a, fast and (computationally) cheap substitute for LOOCV, especially for huge data sets. Decreasing the number of folds (smaller k) may considerably decrease the variance in biasness(smaller data set D/D_t for fitting). Kohavi recommended to take 5-fold or 10-fold cross validation as a better compromise, and suggested for further bias reduction to use a stratification approach or repeated runs.

2.7.2 Bootstrapping

The Bootstrap family was introduced by Efron (2000), and analogous to cross validation it seeks to estimate the expected generalization error. Apart from that, the intrinsic difference between these two techniques are that the bootstrap re-samples the present data at random with replacement, where as cross validation does it without replacement. Suppose we have a training set τ and we wish to fit a model to this data. The basic idea of bootstrapping is to taking data sets randomly with replacement from the training observations, and every sample have similar size i.e, τ, and to repeat this process B(e.g, B=1000) times, producing B bootstrap data sets S_b, where $b = 1, 2, ..., B$. Then the model is refitted to each of the bootstrap data sets, providing a holistic picture of the behavior over the fits. Common methods for estimating the expected prediction error include the LOOCV bootstrap estimate and the 0.632 estimate, which is an extended and improved, in terms of bias reduction, variant of the former. Let the leave-one-out bootstrap estimate be denoted by $\tilde{Err}^1$ and defined like the following,

$$\tilde{Err}^1 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|D^{-i}|}\sum_{b\epsilon D^{-i}} L(y_i, \tilde{f}^{*b}(x_i)) \tag{2.32}$$

where D^{-i} is the set of indices ($|D^{-i}|$), is the number of the bootstrap subsets S_b and b=1,2,...,B that do not involve value i. Bearing in mind that the probability of any observation not being chosen after n samples (with $p = \frac{1}{n}$).

$$P(obs.i \notin S_b) = (1-p)^n = (1-1/n)^n \approx exp^{-1} \approx 0.368 \tag{2.33}$$

Hence, the mean number of particular observations in every bootstrap sample is about $0.632.n$, from which one may infer that leave-one-out bootstrapping has similar characteristics (low variance/high bias) as two-fold cross validation. To alleviate the high bias of the last estimate another bootstrap method was proposed, which is referred to as, 0.632 estimator, (0.632=1-0.368 relates to (29)).

2.8 Measures of Goodness-of-Fit

2.8.1 Receiver Operating Characteristics(ROC)

The ROC graph, originating from signal detection theory[5], is a conceptually simple technique for visualizing, originating and selecting classifiers based on their performance. In general, a classification model(classifiers) is a mapping from instances to predicted continuous (e.g, an estimate of an instances class membership probability) or discrete groups. Let suppose a two group prediction accuracy(binary classification), in which the results of y either are, positive ($y_i = 1$) or negative ($y_i = 0$). For example, if the outcome from a prediction is $\hat{y_i} = 1$ and the actual value is also $y_i = 1$, then it is called a true positive (tp). In table 2.1, the four possible outcomes from a binary classifier are illustrated, from which several common matrices may be derived.

Table 2.1: 2 × 2 confusion matrix.

$\hat{y}$ / y	1	0	Total
1	True +ve	False +ve	P
0	False -ve	True -ve	N
Total	P	N	

$$true - positive - rate = \frac{Positives - Correctly - Classified}{Total - Positive} \tag{2.34}$$

$$false - positive - rate = \frac{Negitives - incorrectly - classified}{Total - Negatives} \tag{2.35}$$

$$Specificity = \frac{True - Negatives}{FalsePositive + TrueNegatives} \tag{2.36}$$

The ROC curve graph plots the probability of detecting true signal(true positive rate on the y-axis) versus a false signal(false positive rate on the x-axis) for an entire range of possible cut points[6], and may also be considered as a tool for visualizing the trade-off between benefits(tp rate) and costs(fp rate). Points in the, north west, of the ROC space reflect higher tp rate and lower fp

[5] One of the first uses was during Word War II for the research of radar waves. Succeeding the attack on Pearl Harbor in 1941, the US army was started new research to improve the prediction of correctly recognized Japanese aircraft from their radar signals.

[6] As the cut points vary, the elements of $\hat{y}$ are different classified. E.g., cut point: 0.5, if $\hat{y} \geq 0.5 \rightarrow \hat{y}$=1, else $\hat{y}$=0

rate and are therefore regarded as desirable [7]. For instance, if the objective is to maximize the classification by choosing an appropriate cut point, one would select a cutoff level that maximizes the correctly and minimizes the incorrectly classified outcomes. The ROC curve, however, is as a quantitative measure of predictiveness of competing models only useful to a limited extent. A common alternative way is to calculate the area under the ROC curve(AUC), which ranges from 0 to 1 and provides a scaler measurement of the classifier′s ability to discriminate. Generally, spoken, $0.7 \leq AUC < 0.8$ is considered as an acceptable and $0.8 \leq AUC$ as an excellent discrimination. For a deeper insight into this topic, such as multi-class ROC graphs or ROC curve averaging (cp. cross validation). In concluding this section, I particularly want to point out that seeking the true model is searching for an impossibility. Rather it is important to seek for a model, which is easily understandable, Parsimonious, appropriate for the situation and which gives a plausible approximation to reality.

[7] Simple random class guessing would produce a diagonal line in the ROC space

Chapter 3

Modified Adaptive LASSO

In this chapter we develop a method for visualization and predictor selection. Application to given research data and simulation study shows that our method outperforms the LASSO, Elastic Net and adaptive lasso in both prediction and variable selection. We consider variables selection technique in which L_1 norm in the penalty part are weighted through sample dependent weights of ridge regression coefficients instead of least square estimates and marginal regression coefficient under certain restrictive conditions(Huang and Zhang,2008), the partial orthogonality condition which does not exist in many situations. These initial estimators of ridge regression coefficients are used as weights by performing relatively better by allowing higher penalty for zero coefficients and lower for non zero coefficient as compared to least squares and marginal regression.

The other advantage of the ridge coefficients is that it never become zero hence it is also used as weights for high dimensional data sets where number of parameters(P) are greater than numbers of observations(n). Moreover, with suitable selection of regularization and control parameters, we demonstrate that the modified estimators do like the oracle approach in predictor selection, it means that it work as well as if the correct model were known.

3.1 Introduction

Predictor selection is a crucial aspect in regression analysis. Usually, a huge amount of explanatory variables are included in the model at opening stage to minimizes potential biasness in the models. To increase predictability and to choose important predictors, statisticians often apply penalization regression procedures. Because these are continuous and also have their theoretical properties are somewhat easy to understand.

Hoerl and Kennard (1970) suggested the ridge regression which estimates the predictor coefficients via an $L_2 - norm$ constrained least-squares criterion. The ridge problem minimize a penalized sum of squares,

$$L(\lambda, \beta) = |y - X\beta|^2 + \lambda|\beta|^2 \tag{3.1}$$

Here $\lambda \geq 0$, is a tuning or condense parameter that controls the model complexity and shrinkage. The co-efficients are shrunk towards zero, but never become zero. It means that all P co-efficients in a ridge fit will be not equal to zero. So that it is not variable selection method. The second drawback is that if there are some correlated predictors in a given multiple regression model, then ridge co-efficients become poorly estimated and exhibit high variance.

The Lasso was proposed by Tibshirani (1996), which is a penalized least squares method by imposing an $L_1 - penalty$ on the regression coefficients.it minimize the equation

$$L(\lambda, \beta) = |y - X\beta|^2 + \lambda|\beta|_1 \tag{3.2}$$

Here the L_2 ridge penalty i.e, $\lambda \sum_{j=1}^{p} \beta_j^2$ is replaced by the $L_1 - penalty$, form as $\lambda \sum_{j=1}^{p} |\beta_j|$. The ridge method has a good prediction performance through a bias-variance trade-off, while the Lasso method encourages both shrinkage and automatic variable selection simultaneously. So that the LASSO has revealed advantages in various scenarios, but it has many restrictions, one of which is the Lasso estimator does not satisfy the oracle properties.

One of the modify Lasso constraint function so that big coefficients reduces less severely, the Smoothly clipped absolute deviation(SCAD) constraint expression of Fan and Li(2005) substitute

$\lambda|\beta|$ to continuous differentiable penalty function. However SCAD suggested criterion is non-convex, so that it is a serious problem with this criterion for optimal solution so that it result to the computational aspect much more difficult.

In case of ($P \gg n$), or if there is a group of correlated variables then Lasso does not provide the satisfactory results(Tibshirani, 1996). Therefore, Zou and Hastie(2005) introduced a new regularization technique called Elastic Net, and its penalty is defined by

$$P_{\lambda_1,\lambda_2}(\beta) = (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \tag{3.3}$$

It is a convex combination of Lasso and ridge penalties. The elastic net gives satisfactory results in these situations, but it has weak oracle properties in some cases. Adaptive Lasso(2006) utilized a weighted constraint of the shape, $\sum_{j=1}^{p} \hat{w}_j|\beta_j|$, in this penalty weights are defined such as $\hat{w}_j = 1/|\beta_j|$, and β_j are ordinary least squares estimators.

This method generates consistent estimates for the interested parameters while maintaining the useful convexity property of the Lasso. Therefore, the adaptive Lasso(2002) has also an oracle characteristics in the view of Fan and Li (2001) and Fan and Peng, (2004). The oracle properties are: the zero components of the true parameters are estimated(shrunk) as 0 with probability tending to 1(also called sparsity property) and the nonzero components have an optimal estimation rate (and is asymptotically normal).

As we already know that the OLS estimators become infeasible in high dimension regression problems, so that adaptive Lasso (Zou, 2006) is not possible to calculate. Huang and Zhang(2008), consider in high dimensional ($P \gg n$) situation a partial orthogonality condition in which the covariates that have 0 coefficients are 0 correlated with the covariate with nonzero coefficient, then marginal regression may be applied to get the preliminary estimator. First the partial orthogonality condition is rarely met in regression analysis. Secondly, even if this assumption exist we cannot apply marginal regression.

Therefore, we suggested ridge coefficients are used as weights instead of OLS estimator in adaptive LASSO(2006), one advantage of ridge coefficients is that it never become zero in high dimensional data analysis, secondly ridge coefficients are consistent and almost unbiased, because ridge coefficient are getting easily whenever we implemented the common OLS technique on an augmented data set.

3.2 Modified Adaptive LASSO

3.2.1 Definition

We assume that a sample data contains, n observations and P covariates. Suppose $y = (y_1, ..., y_n)^T$ and $X = (X_1, ..., X_P)$ be the response vector and the design matrix respectively, here $X_j = (x_{1j}, ..., x_{nj})^T forj = 1, 2, ...P$ are explanatory variables. In change of origin and scale transformation, we may suppose that the dependent variable is centered and the regressors are standardized,

$$\sum_{i=1}^{n} y_i = 0, \sum_{i=1}^{n} x_i = 0, and \sum_{i=1}^{n} x_{ij}^2 = 1 \tag{3.4}$$

for $j = 1, ..., P$

For some constant non-negative λ, we define the modified adaptive Lasso criterion,

$$L(\lambda, \beta) = |y - X\beta|^2 + \lambda_n \sum_{j=1}^{P} w_{ij} |\beta_j| \tag{3.5}$$

Where $w_{ij} = |\tilde{\beta}_j|^{-1}$ are known initial ridge estimators. The value of $L(\lambda, \beta)$ that minimizes the last equation is named modified adaptive Lasso. By allocating a comparatively higher penalty for zero coefficients and, lower value for nonzero coefficients, this modified adaptive Lasso method desires to decrease the estimation biasness and improve variable selection consistency, compared with the standard adaptive Lasso (Zou, 2006).

It is a convex optimization question therefore it does not seriously affected from the so many local optimum values issue, it belong to global minimizer which can be solved easily. This is also an L_1 penalization technique. We can employ the prevailing productive algorithms to compute the Lasso, follow the same, we can also get the modified adaptive estimates.

To find an optimal tuning parameter, λ a major topic in practice. Assume that if we utilize $\tilde{\beta}$

(OLS) to build the weights for adaptive lasso; furthermore, to obtain an optimal value of λ. We can implement 10-fold cross-validation, or BIC approach to get the best value of λ for adaptive lasso. In our approach, we replace $\tilde{\beta}$ (OLS) with ridge estimators which are consistent under some condition. Thus we can handle similarly the second shrinkage parameter and run two-dimensional cross-validation to get a best pair of $\{\tilde{\beta}(ridge), \lambda\}$. We worked to apply $\tilde{\beta}$ (ridge) instead of OLS estimator, from the best ridge regression fit, because it is more stable than $\tilde{\beta}$ (OLS) particularly in $P \gg n$ case.

3.2.2 Solution

It is well known that minimizing the problem (3.5) equivalent to Lasso-type optimization problem. This fact implies that the adaptive Lasso type penalty also enjoys the computational advantage of the Lasso. The other advantage is that the modified adaptive Lasso, can be obtained from adaptive Lasso (2006) on augmented data. Because the ridge regression estimates can be derived by applying ordinary least squares regression(OLS) on an augmented data set. This can be proved by using lemma1.

Lemma 1: Given data set (y, X) and λ be the regularization parameter, define an augmented data set (y^*, X^*) by

$$X^* = \begin{pmatrix} X \\ \sqrt{\lambda} I \end{pmatrix} \tag{3.6}$$

$$y^* = \begin{pmatrix} y \\ 0 \end{pmatrix} \tag{3.7}$$

where I is $P \times P$ identity matrix, and y^* is augmented with P zero rows. Now the ordinary least squares estimates is defined by,

$$L(\beta^*) = (y^* - X^*\beta^*)^T (y^* - X^*\beta^*) \tag{3.8}$$

where β^* is OLS estimator based on augmented data. Solving just by simple algebra after assume orthonormal design matrix X, we have

$$L(\beta^*) = (y^T y - 2y^T X\beta + \beta^T\beta + \lambda\beta^T\beta) \tag{3.9}$$

Differentiating with respect to β, and equate to zero, we get

$$\widetilde{\beta} = \frac{y^T X}{1+\lambda}, \tag{3.10}$$

Where $y^T X$ is simple ordinary least squares estimates based on original data set.
Clearly, lemma 1 shows that, $\widetilde{\beta}$ is OLS estimators based on augmented data set, is equivalent to ridge estimators. The proof is given an Appendix A.1.
Theorem.1: Given the data in Lemma1, let $\theta = \lambda_{1n}(1+\lambda_{2n})^1/\gamma$ and $\beta^* = \beta$ then modified adaptive Lasso criterion can be written as,

$$L(\theta, \beta^*) = ||y^* - X^*\beta^*||^2 + \theta \sum_{j=1}^{P} w_{ij}^* |\beta_j^*| \tag{3.11}$$

Let

$$\beta^* = argmin_{\beta^*}\{L(\theta, \beta^*)\} \tag{3.12}$$

Then

$$\widetilde{\beta} = (1+\lambda_{2n})\beta^* \tag{3.13}$$

The proof is simple and given in Appendix A.2. Theorem 1 says that we can obtained the Adaptive Lasso (Zou, 2006) problem into an equivalent to modified Adaptive Lasso problem on augmented data. This essential property overcomes the limitations of the Adaptive Lasso (Zou, 2006) in high dimensional data, $(P \gg n)$ situation.
The modified adaptive Lasso estimates β^* are defined by,

$$\beta^* = \frac{\widetilde{\beta}}{1+\lambda_{2n}} \tag{3.14}$$

$$\beta^*(modifiedA - LASSO) = \frac{\widetilde{\beta}}{1+\lambda_{2n}} \tag{3.15}$$

with standard adaptive Lasso $\widetilde{\beta}$ (H.Zou,2006).

Hence, modified Adaptive Lasso coefficient is a rescaled adaptive Lasso(H.Zou,2006) coefficient. Therefore the everyone delightful property of the adaptive Lasso(2006) hold for modified adaptive Lasso procedure too.

Let consider the orthogonal design matrix X, i.e, $X^T X = 1$, the solution of modified method for unknown coefficients β, if the tuning parameters(λ_1, λ_2) are properly selected, is form as:

$$\hat{\beta}^{ML} = \hat{\beta_j}^{OLS} - \frac{\lambda_1}{2}\hat{w_j}^* sign(\hat{\beta_j}^{OLS}) \tag{3.16}$$

In above equation, $\hat{\beta}^{ML}$ is represents the modified LASSO estimator, for multivariate case the OLS estimator is defined like $\hat{\beta_j}^{OLS} = X^T y$ and weights for modified procedure is $\hat{w_j}^*$. The complete proof is given an appendix A.3.

An efficient and a stylized type of the Stagewise procedure that utilize a simple mathematical formula to accelerate the computations is Least Angle Regression (LARS) proposed by Efron (2004). They showed that, beginning from 0, the Lasso solution paths spread piecewise linearly in a predictable style. They advocated a fresh method named least angle regressions(LARS) to explain the entire LASSO solution path efficiently by working in similar layout of computations as a alone OLS fit. This modified predictor selection approach can also be solved by applying LARS algorithm. The computation of modified adaptive Lasso is given in the following algorithm.

3.2.3 LARS Algorithm for Modified Approach

The LARS algorithm has been introduced by Efron et.al, (2004) to explain the computational problems in penalized least squares. The application of this algorithm in our method is given below,

We first define

$$X_j^{**} = \frac{X_j}{\widetilde{w}_{nj}} \tag{3.17}$$

For j=1,2,...,P. Solve the last equation for the LASSO problem for all λ_n, we get

$$\tilde{\beta}^{**} = argmin_{\beta}||y - \sum_{j=1}^{p} X_j^{**}\beta_j||^2 + \lambda_n \sum_{j=1}^{p} ||\beta_j||_1 \tag{3.18}$$

The output is,

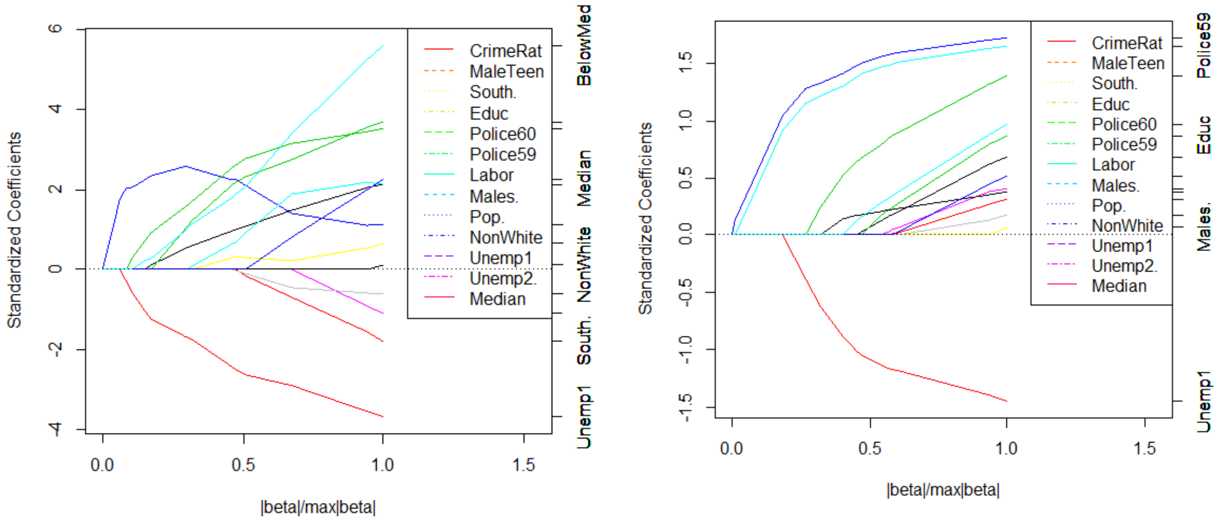
$$\tilde{\beta}^{***} = \frac{\tilde{\beta}^{**}}{\tilde{w}_{nj}} \tag{3.19}$$

This coefficient path algorithm model the modified adaptive Lasso a nice-looking method for real data application.

The solution paths of modified adaptive Lasso and elastic net for Crime data set are shown in figure 3.1. The values of shrinkage factor plotted on X-axis are chosen by cross validation. The modified LASSO coefficients are plotted on Y-axis. The two estimates are linear multiple sub-functions, it is an important characteristic of the LARS formula. The optimum value is chosen between the 0 and 1 of factor values. For instance, if we choose 0.5, then the coefficients right of 0.5 hit to zero by modified method. But from figure 3.1(a) we can see that, the modified adaptive LASSO coefficient profile is not stable, in opposed, the elastic net has more well ordered solution pathways.

3.3 Properties of Modified Adaptive Method

In this section we discuss some useful properties of penalized regression methods.



(a)

(b)

Figure 3.1: (a) Modified LASSO and (b) elastic net solution paths.

3.3.1 Selection of Model Complexity Parameters

Currently, we want to explore the selection of shrinkage parameters in modified adaptive Lasso technique. Assume that if we utilize the consistent $\widetilde{\beta}$ (Ridge) to generates the weights for modified procedure, we then wish to achieve a best pair of tuning parameters one for ridge regression fit to be used its coefficients as weights and other for adaptive Lasso. Let λ_1 and λ_2 are corresponding regularization parameters of ridge and adaptive respectively. There are well-regularized methods for selecting similar tuning parameters. If entire training data is accessible, 10-fold cross-validation is a well liked procedure for estimating the prediction error and evaluating the efficiency of separate models, and we want to apply this technique in my thesis work. However, in this modified adaptive method two shrinkage parameters are observed, therefore we require to perform cross-validation on a two-dimensional side. Customarily we initially fit ridge regression to compute weights for modified adaptive Lasso method, so for it we find a best value for λ_1 over a grid of values, say (0.1,1,length=100). Then, for already find value of λ_1 ,we will implement cross-validation procedure just same as discussed along with the algorithm LARS-EN, and try to determine for the optimal λ_2.

3.3.2 The Grouping Effect

Sometimes the explanatory variables associated to some pre-defined group, for instance genes that reported to the same biological pathway, or some of discrete variables for displaying the levels of a categorical independent variable. Such problem has been extensively addressed in literature. For example, Hastie et.al (2005) and Diaz and Uriarte (2003), uses principal component analysis. Hastie et.al(2003) applies supervised learning approaches, Tree harvesting, and some other procedures to choose classes of predictive genes built by hierarchical clustering. A research by Segal and Conklin (2003) proposed to utilize the regularized regression methods to detect the grouped genes. In these situations it may preferable to diminish and choose the variables of a group or a class together. When there is a set of such predictors and the pairwise relationships among them is extreme, after that LASSO perform poorly. Zou, et.al (2005) proposed a new method which perform better than Lasso in that situation. We also evaluated the performance of modified adaptive Lasso in that situation.

3.3.3 Standard Error Formula

We use Fin and Li (2001) proposed LQA (Local quadratic approximation) sandwich formula for solving the covariance matrix of constraint method estimates for coefficients which are not zero. Consider the LQA formula for modified adaptive Lasso penalty, for nonzero coefficients β_j,

$$|\beta_j|\tilde{w}_j \approx |\beta_{j0}|\tilde{w}_j + \tfrac{\tilde{w}_j}{2|\beta_{j0}|}(\beta_j^2 - \beta_{j0}^2)$$

Where $\tilde{w}_j$ are weights of ridge coefficients. Consider that the primary l elements of β are not equal to 0.

$$\sum(\beta) = diag(\frac{\tilde{w}_1}{|\beta_1|}, ..., \frac{\tilde{w}_l}{|\beta_l|}) \tag{3.20}$$

Let X_l denote the first l columns of design matrix. Fin and Li (2001) solved the adaptive Lasso estimates by iteratively computing the ridge regression, we want to use just like it by solved modified adaptive Lasso,

$$(\beta_1, ..., \beta_l)^T = (X_l^T X_l + \lambda_n \sum \beta_0)^{-1} X_l^T y \tag{3.21}$$

It can be use to estimate the matrix of covariances for the nonzero elements of β estimates i.e, β_n^{**} of modified method. Let

$$\kappa_n^* = (J\colon \tilde{\beta}_n^{**} \neq 0) \tag{3.22}$$

then,

$$\widetilde{cov}(\tilde{\beta}_{\kappa_n^*}^{**}) = \sigma^2\{(X_{\kappa_n^*}^T X_{\kappa_n^*} + \lambda_n \sum(\tilde{\beta}_{\kappa_n^*}^{**}))^{-1} X_{\kappa_n^*}^T X_{\kappa_n^*}(X_{\kappa_n^*}^T X_{\kappa_n^*} + \lambda_n \sum(\tilde{\beta}_{\kappa_n^*}^{**}))^{-1}\} \tag{3.23}$$

If σ^2 is unknown, then we can use its estimate from full model. The variables for which $\tilde{\beta}_n^{**} = 0$ then the standard error estimates are also zero(Tibshirani 1996;).

3.3.4 Oracle Properties

The Oracle properties means that, the zero coefficient of true parameters are estimated as 0 with probability approaching to 1, (it is also called sparsity property) and the non-zero coefficient have an optimal estimation rate, (and is asymptotically normal). According to theorem(1), the modified adaptive Lasso coefficient is a rescaled Adaptive Lasso(2006) coefficient. Hence, both the oracle properties of the Adaptive LASSO(Zou,2006) also hold for the $modified$ adaptive LASSO. Through simulation study, we show that the modified method works very good just like a oracle estimator when compared with other regularization and shrinkage methods too. Recall that,
(a) $y = x_i\beta^* + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$
(b) $\frac{1}{n}X^TX \rightarrow D$, where D is positive definite matrix. Now assume that $\Theta = (1, ..., p_0)$. Let

$$D = \begin{bmatrix} D11 & D12 \\ D21 & D22 \end{bmatrix}$$

Where $D11$ is a $p_0 \times p_0$ matrix of predictors that associated with non-zero coefficients.
We take the modified method estimates,$\tilde{\beta^n}$ are as under,

$$\tilde{\beta^n} = argmin_\beta ||y - \sum_{j=1}^{p} x_j\beta_j||^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \tag{3.24}$$

where λ_n depend on n.Let

$$\Theta_n = (j : \tilde{\beta_j^n} \neq 0) \tag{3.25}$$

The proposed adaptive LASSO will be selected variables consistently if and only if

(1) $lim_n P(\Theta_n = \Theta) = 1$.

(2) For asymptotic normality; $\sqrt{n}(\tilde{\beta_\Theta^{(n)}} - \beta_\Theta) \to N(0, \sigma^2 \times D11^{-1})$

This indicates that the L_1 penalty is at least as better as compared to any other, and work like oracle procedure. The proof and several remarks are given by Zou (2006).

3.4 A Simulation Study

The basic aim of this simulation study is to show that the modified adaptive Lasso dominates the other penalized regression methods with regard to prediction accuracy and predictors selection in different scenarios. We also study modified adaptive Lasso in low-dimension($P < n$) and high-dimensional ($P \gg n$) surroundings. For each method, we use five-fold or ten-fold cross validation to estimate the model complexity parameters. We simulate data from true model,

$$y = X^T\beta + \sigma\epsilon_i \tag{3.26}$$

where $\epsilon_i \sim N(0, 1)$.

Example 4.1 We simulated 1000 data sets containing of 100 observations and eight predictors. We let $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. The components of X and ϵ_i are standard normal. The pairwise correlations between X_i and X_j is set to be $corr(X_i, X_j) = 0.5^{|i-j|}$. The average number of zero's estimates is also reported in all tables. The column marked Correct represents the mean numbers of zero coefficients correctly estimated as 0, and the column marked Incorrect outline the average of estimates erroneously set to 0.

Example 4.2 Example 4.2 is identical to model 4.1, except that $\beta_j = 0.8$ for all j=1,...,P.

Example 4.3 This example is the same as example 4.1, except that $corr(X_i, X_j) = 0.9^{|i-j|}$, and also consider low correlation between predictors i.e, $corr(X_i, X_j) = 0.15^{|i-j|}$.

Example 4.4(High-dimensional dataset) We simulated 1000 data sets consisting of 100 observations and 200 explanatory variables. We also set $\beta = \{(3, ..., 3)_5, (1.5, ..., 1.5)_5, (2, ..., 2)_5, (0, ..., 0)_{185}$ and $\sigma = 5$. The pairwise correlations between explanatory variables are same as given in example 4.1.

Example 4.5(Grouping effect) We simulated 1000 data sets consisting of 200 observations and 40 predictors. We choose $\beta = \{(3, ..., 3)_{15}, (0, ...0)_{25}\}$ and $\sigma = 3$. The independent variables are generated as follows:
$Z_k \sim N(0, 1)$ for k=1,2,3; and $\epsilon_i^x \sim N(0, 0.01)$ for i=1,2,...,15.
$X_i = Z_1 + \epsilon_i^x$, for i=1,...,5.
$X_i = Z_2 + \epsilon_i^x$, for i=6,...,10.
$X_i = Z_3 + \epsilon_i^x$, for i=11,...,15.
The remaining predictors are independently and identically distributed with mean 0 and standard deviation 1 i.e, $X_i \sim N(0, 1)$ for i=16,...,40. In this scenario we have constructed three identically vital groups, and inside everyone group there are five components. The fourth group has 25 independently and identically noise features. A good method would be that which set entirely the 15 true predictors and exclude the rest of 25 coefficients which are considered as zero.

Example 4.6 Example 6 is the same as example 5, except that P=400.

Example 4.7(Robust regression) In this robust regression case we consider 1000 simulation replications, n=100 and $\sigma = 3$. We simulated datasets from the model

$$Y = X^T \beta + \sigma \epsilon, \tag{3.27}$$

Table 3.1: Median prediction errors for the simulated example 4.1 of six concerned procedures found on 1000 replications. The bias is averaged for only non-zero coefficients.

			Avg.No. of 0 Coefficients	
Method	Med.PMSE	Avg.Bias	Correct	Incorrect
Ridge	10.971	(-2.26,-0.78,-1.38)	...	...
LASSO	11.212	(0.49,-1.31,-0.69)	3.4	0.09
Elastic Net	10.775	(-0.78,-0.24,-1.17)	2.08	0.03
Adaptive LASSO	11.092	(0.59,-1.5,0.24)	4	0.16
Proposed	9.720	(-0.38, 0.20,0.13)	4	0
Oracle	9.313	(-0.242,0.43,0.063)	5	0

The matrices X and β are similar as specified in example 4.1. The only difference is to generate ϵ from standard gaussian distribution including 10 % outliers from the standard Cauchy distribution. We also know that the least squares estimates are not robust against extreme observations. Therefore, we take the outlier-resistent loss functions, for instance, absolute loss function, Huber's ψ function and bi-square function. Hence, we minimize,

$$\sum_{i=1}^{n} \psi(|y_i - x_i\beta|) + \sum_{j=1}^{P} p_\lambda(|\beta_j|) \tag{3.28}$$

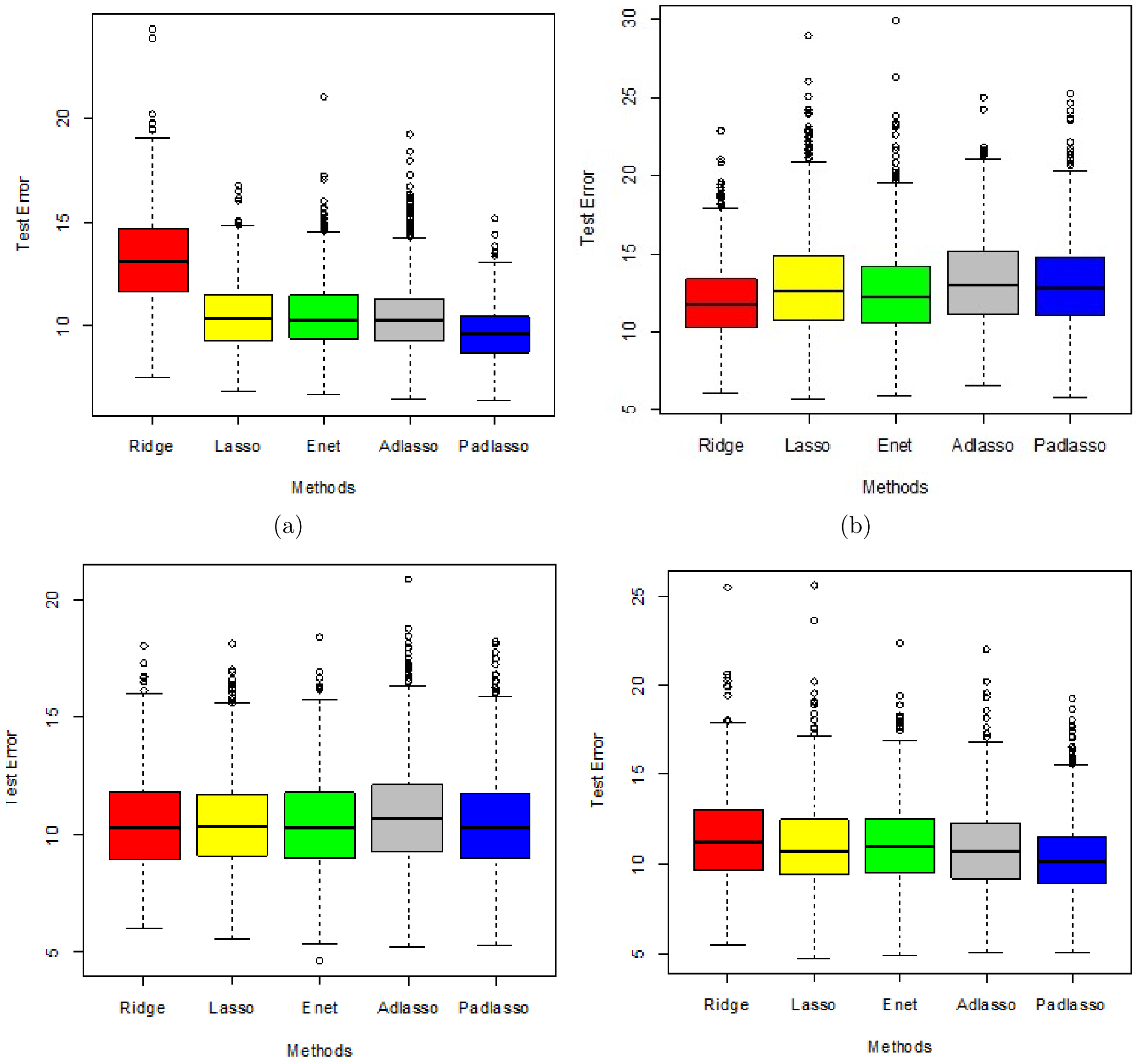
w.r.t β. This ψ function gives regularized robust estimator for β.

Table 3.1 and figure 3.2(a) summarize the median of prediction mean-squared errors results. We observe that the *modified adaptive Lasso* has good performance (in example 4.1) than other penalized methods with regards to both prediction accuracy and variables selection. The prediction error of modified adaptive Lasso is about 12% lower than that of adaptive Lasso (Zou, 2006), and 10%, 13%, and 11% lower than those of Elastic net, Lasso(Tibshirani,1996)and Ridge regression respectively. It can also be seen that the average bias for nonzero coefficients is also minimum of modified adaptive Lasso than other regularization methods. Now consider the variable selection, from table 3.1 we can see that the modified adaptive Lasso performs best and reduce model complexity correctly. The adaptive Lasso(Zou,2006) and modified adaptive performs almost similar (both selected 4 variables) in variables selection but modified adaptive Lasso is zero error of average number of "0" coefficients estimated incorrectly, and sometime the adaptive Lasso(H.Zou,2006)

Table 3.2: Simulation results for example 4.2; and bootstrapp estimates of standard errors given in parentheses, by using bootsrapp with B=500 resamplings on the 1000 test mean squared errors.

	Results	
Method	Med.PMSE	Avg.No.of 0 Coefficients
Ridge	11.769(0.083)	0
LASSO	12.605(0.101)	1.58
Elastic Net	12.214(0.09)	1.32
Adaptive LASSO	12.941(0.087)	3.16
Proposed	12.839(0.094)	1.802
Oracle	11.364(0.079)	0

is selected "0" coefficient incorrectly, here is 0.16 times (16 times in 100). The other methods perform poorly in variables selection, whereas ridge regression reduces only model error and select all variables every time. Hence, the performance of modified adaptive Lasso is believed to be as close as that of the oracle estimator with compared to the other methods in example 4.1. Table 3.2 and figure 3.2(b) shows that the ridge regression is significantly more accurate than other methods in case of many small non zero effects. In this example the adaptive Lasso(Zou, 2006) has performed very poor in terms of both, prediction accuracy and variables selection. The bootstrap standard errors of the prediction errors with B=500 re-samplings on the 1000 test errors are given in parentheses. The smallest standard error is also corresponding to ridge regression, which shows dominance in scenario of many small non-zero coefficients if the true model is given. Table 3.3 and its corresponding box plots are given in figure 3.2 (c) and (d), and summarized the simulation results of example 4.3 for low and high pair wise correlations between predictors. Observing table 3.3, it can be noticed that the modified adaptive Lasso comparatively surpasses the other variables selection approaches. In variables selection again the other methods perform very poor and modified adaptive Lasso enjoys the oracle properties. Similarly we can also say, the modified adaptive method does not damage through excessive correlation among predictors. The findings also shows that the ridge regression is minimum prediction error when highly correlated predictors are included in the model.



(a)

(b)

(c)

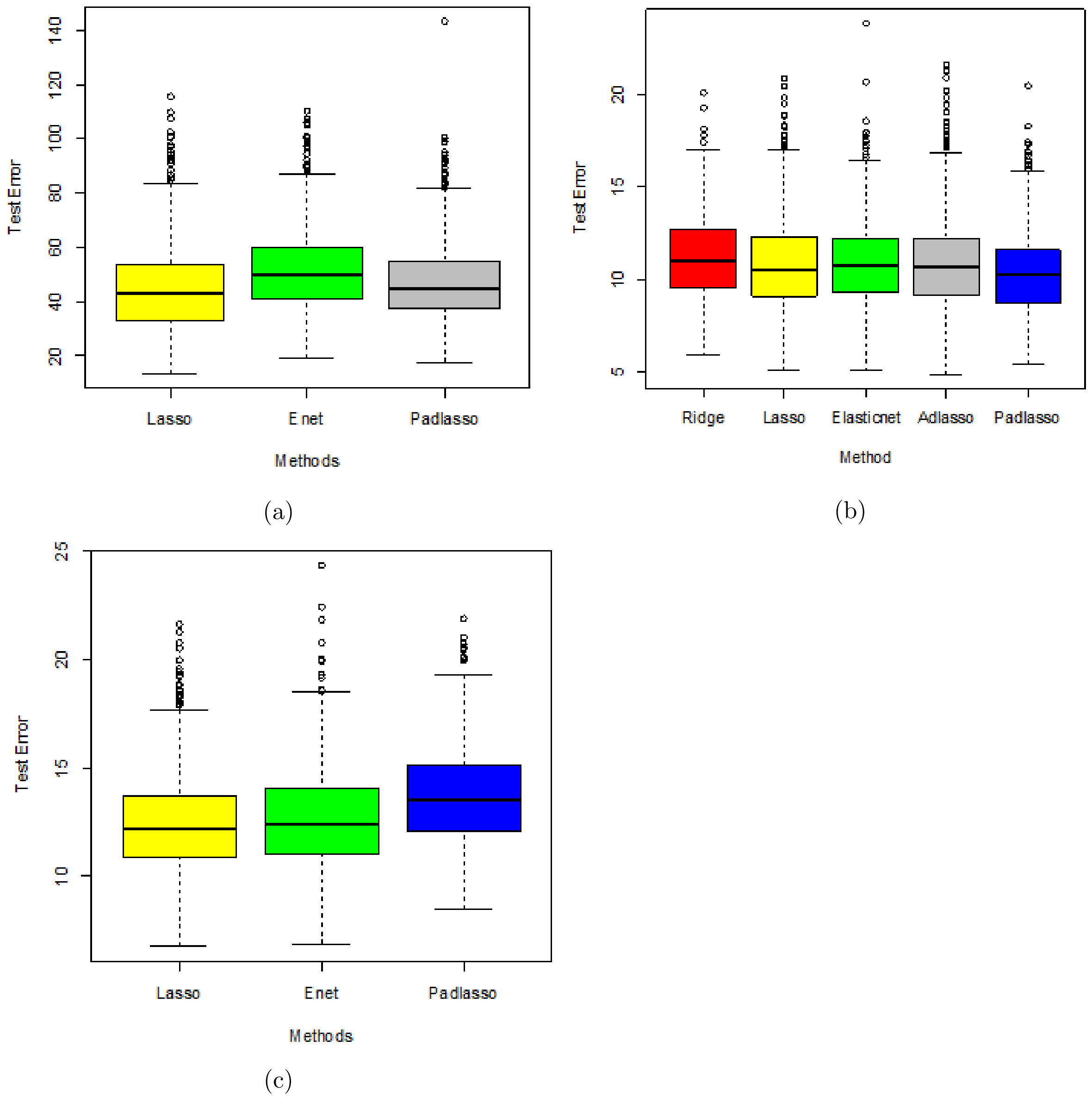
(d)

Figure 3.2: Boxplots, of comparing the prediction accuracy of ridge regression, LASSO, elastic net, adaptive LASSO and modified adaptive LASSO in all four examples: (a) example 4.1; (b) example 4.2; (c) and (d) example 4.3.

Table 3.3: Simulation results of example 4.3, Comparing the Median Prediction Mean Squarred Errors Based on 1000 Replications

		Avg.No. of 0 Coefficients	
Method	Med.PMSE	Correct	Incorrect
$n = 100, \rho = 0.9$			
Ridge	10.259(0.065)	...	...
LASSO	10.353(0.061)	3.032	0.219
Elastic Net	10.298(0.065)	2.431	0.161
Adaptive LASSO	10.664(0.074)	3.583	0.835
Proposed	10.263(0.065)	4.597	0.319
Oracle	9.916(0.06)	5	0
$n = 100, \rho = 0.15$			
Ridge	11.258(0.081)	...	....
LASSO	10.779(0.077)	2.541	0.013
Elastic net	10.942(0.069)	2.231	0
Adaptive LASSO	10.703(0.072)	3.807	0.079
Proposed	10.131(0.071)	4.485	0
Oracle	9.835 (0.063)	5	0

The summary of simulation results for high-dimensional($P \gg n$) dataset, given in example 4.4, is depicted in table 3.4 and figure 3.3(a). In given example we consider 200 variables and only 100 observations. Clearly, we can see that the adaptive LASSO(Zou,2006) is absent in the table because in this situation ($P \gg n$) it is nontrivial to calculate reliable estimates for weights using in it. The findings shows that the ridge regression has a very poor performance, both in prediction accuracy and variable selection. From table 3.4, it can be seen that the LASSO reduces the model error, but in variable selection it performed poor against proposed method, therefore, the modified adaptive LASSO produces more sparse solutions and acts just like oracle estimator. The bootstrap standard error with B=500 based on 1000 test errors of the modified adaptive Lasso is also minimum than others approaches.



(a)

(b)

(c)

Figure 3.3: Boxplots, of comparing the test errors of ridge regression, LASSO, elastic net, adaptive LASSO and modified adaptive LASSO in simulation examples: (a) example 4.4; (b) example 4.5; and (c) example 4.6.

Table 3.4: Simulation outcomes for high-dimensional data set, simulated in example 4.4.

		Avg.No. of 0 Coefficients	
Method	Med.PMSE	Correct	Incorrect
Ridge	403.988(0.53)	...	...
LASSO	42.85(0.503)	167.10	2.01
Elastic Net	49.947(0.482)	162.04	2.39
Proposed	44.817(0.478)	178.91	0.25
Oracle	40.285(0.368)	185	0

In table 3.5 and its corresponding box plots are given in figure 3.3(b) and (c), shows the simulation results of the ”grouped variables” situation (in examples 4.5 and 4.6) for P=40, 400 and n=200. First we can see that modified adaptive weighted technique command over the other methods by a good margin. The suggested method has prediction error about 7 % lower than those of adaptive LASSO(2006), elastic net(2005), LASSO(1996) and 26% lower than ridge regression. Now, in terms of variables selection the adaptive Lasso (Zou, 2006) perform very poor, i.e for P=40 it select 25 variables correctly as 0, which is exactly the same as oracle estimator, but also set about 12 variables to zero incorrectly. It can also be seen that the achievements of modified adaptive technique and elastic net are almost same, both did not estimate as 0 coefficients incorrectly, but in terms of correctly estimate 0 coefficient, modified adaptive Lasso selected about 24 variables and elastic net select 19 variables. Therefore, the suggested method is anticipated to be as better and working just like an oracle estimator in the grouped variables situation. In case of, high-dimension data, in term of prediction accuracy these methods, except ridge regression, almost similar, but again in variables selection the modified adaptive Lasso outperforms the other procedures. Clearly, the standard error(*bootstrap*) of modified adaptive Lasso is minimum than other regularization procedures. The simulation findings of robust regression example(i.e, 4.7) are summarized in table 3.7. The detailed results are also similar to those in example 1. The proposed method gives the lowest prediction error and work just like as oracle in model selection performance. The numerical results in parentheses are the corresponding median absolute deviations(MAD's), evaluated by applying the bootstrap process with B=500, re-samplings on the 1000 simulated test errors. Clearly, we can see that our method has minimum bootstrap median absolute deviation. Table 3.6 presents

Table 3.5: Median mean-squared errors for the simulated models for examples 4.5 and 4.6; and bootstrapp estimates of standard errors given in parentheses, by using bootsrapp with B=500 re-samplings on the 1000 simulated test mean squared errors.

		Avg.No. of 0 Coefficients	
Method	Med.PMSE	Correct	Incorrect
$P = 40, n = 200$			
Ridge	13.126(0.071)	...	...
LASSO	10.375(0.051)	19.62	6.69
Elastic Net	10.34(0.054)	19.22	0
Adaptive LASSO	10.256(0.058)	25	11
Proposed	9.627(0.04)	24.36	0
Oracle	9.741(0.047)	25	0
$P = 400, n = 200$			
Ridge	320.485(1.278)	...	...
LASSO	12.206(0.074)	365.52	6.69
Elastic net	12.238(0.077)	368.84	0
Proposed	13.271(0.063)	382.71	0
Oracle	9.677(0.047)	385	0

Table 3.6: Standard deviations of estimators for the linear rgression model with n=100 and $\sigma = 2$

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
proposed	0.165	0.284 (0.043)	0.167	0.264 (0.035)	0.143	0.27 (0.046)
A.LASSO	0.173	0.264 (0.049)	0.16	0.261 (0.05)	0.168	0.281 (0.065)

Table 3.7: Simulation Results for the Robust Linear Model given in example 4.7

		Avg.No. of 0 Coefficients	
Method	Med.PMSE	Correct	Incorrect
Ridge	21.208(1.535)	...	...
LASSO	18.725(1.216)	2.439	0.10
Elastic Net	18.838 (0.97)	2.92	0.07
Adaptive Lasso	19.44 (1.313)	3.85	0.52
Proposed	16.261(0.82)	4.74	0.19
Oracle	13.456(0.193)	5	0

the results of standard errors calculated by using the sandwich formula for non-zero coefficients. The true standard error denoted by SD in table 6 can be calculated to find the median absolute deviation(MAD) divided by 0.6745 of the 100 estimated coefficients in the 100 simulations. The median of the 100 estimated standard deviations finding through sandwich formula, denoted by SD_m. In table 6 SD_{mad} represents the mean absolute deviation of 100 estimated standard errors(SD_m's) divided by 0.6745. The results shows that the new method is also perform well by using sandwich formula.

3.5 Real Data Example

The data set for this example is the Crime data come from a study between crime rates and demographical data for 47 U.S. states in 1960. The data was originally collected by FBI and the government agencies. The data file contains 14 variables and 47 observations with one for each state in the U.S. The project describe the crime rates and other demographical factors which affect crime rates. Here we want: 1. Do demographical characteristics have impacts on crime rates? 2. If there is relationship between crime rates and demographical data, which factor contributes most to the crime rates?

3.5.1 Data Description

The 14 variables can be described as:

CrimeRat: Numbers of offenses reported to police per million population, this is the dependent variable.

MaleTeen: The total figure of males having age between 14-24 per 1000 population.

South.: Categorical predictor for Southern states of US(0=No, 1=Yes).

Educ: Average number of schooling years x 10 for students of age 25 or above.

Police60: 1960 per capita expenditure invested on police department by respective US state and local government.

Police59: 1959 per capita expenditure invested on police department by respective US state and local government.

Labor: Labor force involvement rate per 1000 civilian males living in urban areas having age 14-24.

Males.: The ratio of males people per 1000 females.

Pop.: US state size of population in 100 thousands.

NonWhite: The totality of non-whites guys per 1000 population.

Unemp1: Unemployment rate of males living in urban area per 1000 of age 14-24.

Unemp2.: Unemployment rate of males living in urban area per 1000 of age 35-39.

Median: Median value of transferable commodities and property or family income in tens of $.

BelowMed: The totality of households per 1000 earnings below 1/2 the median income.

The response variable y is "CrimeRat" and the other predictors are MaleTeen, South., Educ, Police60, Police59,Labor, Males., Pop., NonWhite, Unemp1, Unemp2., Median, and BelowMed.

3.5.2 Data Analysis

The data set contains many variables, therefore, first we check the multi-collinearity between pairs of variables because this issue affect different penalized regression methods and some method do not suffer very much due to this issue. The correlation matrix between predictors in table 2.8, shows some strong correlations between many explanatory variables. For example, Police60 and Police59 show a strong positive correlation (i.e 0.944). The other predictors are Median and police59, Median and MaleTeen, Educ and South. , and many others have strong correlations.

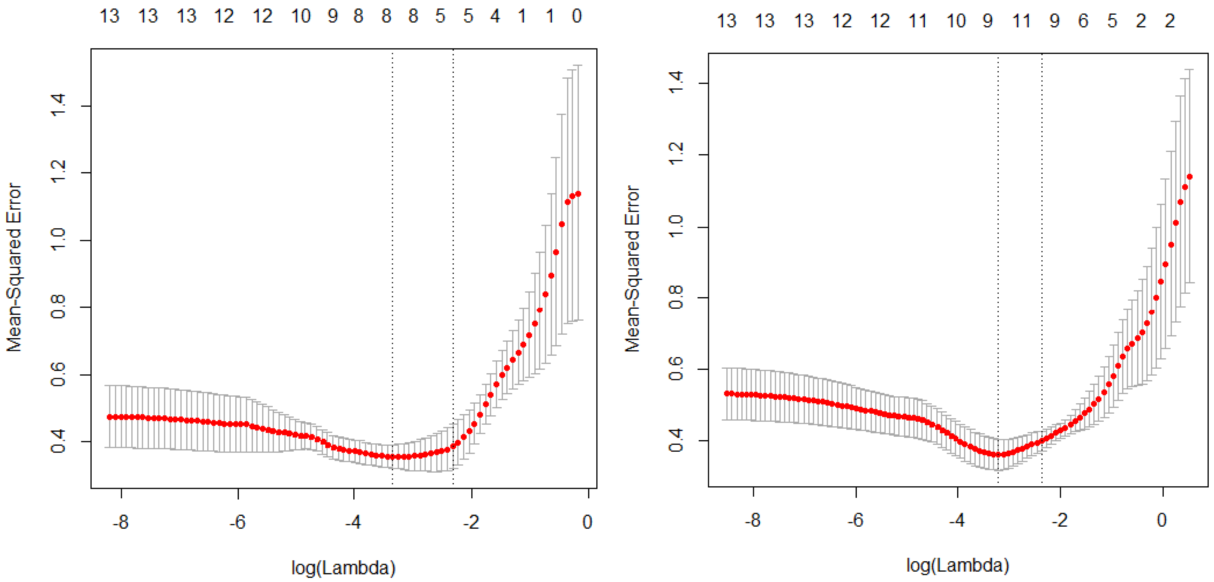
Table 3.8: Correlations of predictors in the Crime data set

Variable	MaleTeen	South.	Educ	Police60	Police59	Labor	Males.	Pop.	NonWhite	Unemp1	Unemp2.	Median
MaleTeen	1.000											
South.	0.584	1.000										
Educ	−0.530	-0.703	1.000									
Police60	-0.506	-0.373	0.483	1.000								
Police59	-0.513	-0.376	0.499	0.994	1.000							
Labor	-0.161	-0.505	0.561	0.121	0.106	1.000						
Males.	-0.029	-0.315	0.437	0.034	0.023	0.514	1.000					
Pop.	-0.281	-0.050	-0.017	0.526	0.514	-0.124	-0.411	1.000				
NonWhite	0.593	0.767	-0.665	-0.214	-0.219	-0.341	-0.327	0.095	1.000			
Unemp1	-0.224	-0.172	0.018	-0.044	-0.052	-0.229	0.352	-0.038	-0.156	1.000		
Unemp2.	-0.245	0.072	-0.216	0.185	0.169	-0.421	-0.019	0.270	0.081	0.746	1.000	
Median	-0.670	-0.637	0.736	0.787	0.794	0.295	0.180	0.308	-0.590	0.045	0.092	1.000
BelowMed	0.639	0.737	-0.769	-0.631	-0.648	-0.270	-0.167	-0.126	0.677	-0.064	0.016	-0.884

Table 3.9: Coefficients estimates and prediction errors, for various shrinkage and variables selection procedures implemented to the Crime statistics. The dashed entrees correspond to predictors omitted.

Variable	Ridge	LASSO	Elastic net	Adaptive LASSO	Modified A.LASSO
Intercept	0.018	0.021	0.023	0.049	0.037
MaleTeen	0.303	0.159	0.223	0.265	0.239
South.	0.130	...	0.043	...	...
Educ	0.217	...	0.139	0.277	...
Police60	0.497	0.890	0.775	1.577	1.001
Police59	0.319	...	0.134	-0.484	...
Labor	0.110	...	...	...	...
Males.	0.176	0.135	0.161	0.034	0.079
Pop.	0.017	...	0.017	...	...
NonWhite	0.131	0.098	0.114	...	...
Unemp1	-0.058	...	...	...	...
Unemp2.	0.187	...	...	...	...
Median	-0.003	...	...	...	...
BelowMed	0.201	0.143	0.241	0.518	0.202
Test Error	0.61	0.631	0.615	0.65	0.606

Before going to direct analysis, first we standardize the explanatory variables to have variance 1, then we randomly allocate the data set into a training set of size 32 observations, and a validation set of size 15. Table 3.9 summarized the coefficients estimates determined applying various predictor selection and shrinkage methods. They are Ridge regression, Lasso, Elastic net, adaptive Lasso (Zou,2006) and modified adaptive Lasso. Each method has a tuning or a complexity parameter(s), and this was selected to minimize an estimate of prediction error on 10-fold cross-validation. This validation approach is employed on training set, hence choosing the complexity parameter(s) is concern of the training process. Figures 3.4 and 3.5, gives the relationship between $log(\lambda)$, choosing by using 10-fold cross-validation, and MSE. The integer numbers above of plots shows the number of estimators or predictors non-zero in the model. The left line gives the smallest MSE with number of predictors present in the model and the right line gives the smallest MSE with number of variables in the model. We can therefore choose any value of $lambda(\lambda)$ between the left line and right line which gives optimum results. The considered learning method is fit for a set of values of the complexity parameter-to nine-tenth of the data and test error is evaluated on the

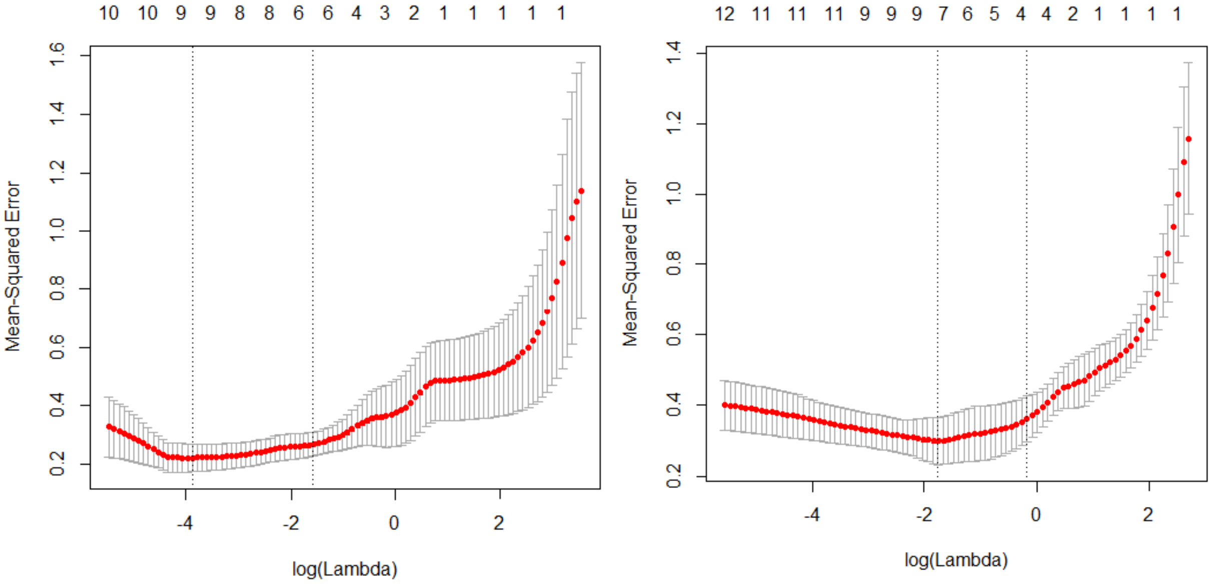


(a)

(b)

Figure 3.4: MSE plots and the number of Variables in the model as a function of $log(\lambda)$ for the 10-fold cross validation for the (a)LASSO regression (b)Elastic Net.

remaining one-tenth. This is done in turn for every one-tenth of the data, and 10 prediction error estimates are averaged. The test set is use to judge the prediction performance of the selected learning method. Table 3.9 shows that LASSO chose to use the five predictors, MaleTeen, Police60, Males., NonWhite, and BelowMed. Elastic net select nine and adaptive Lasso six variables. The $modified\ adaptive\ Lasso$ (suggested) produces more sparse solution, and choose four variables, they are MaleTeen, Police60, Males., and BelowMed corresponding to smallest test error which has given in last row of table 3.9. One main feature of modified adaptive Lasso method is that the four variables selected by it, these four variables are also selected by the other three learning methods. As there is a some more correlations among features the prediction error of ridge regression (i.e, 0.61) is also minimum but not from proposed method.



(a)

(b)

Figure 3.5: MSE plots and the number of Variables in the model as a function of $log(\lambda)$ for the 10-fold cross validation for the (a)Adaptive LASSO penalized regression (b)Modified Adaptive LASSO penalized regression.

Chapter 4

Assessing the Performance of Various Penalized Regression Methods Under Heteroscedasticity

In this chapter our objective is to evaluate the performance of different shrinkage and variables selection procedures when error term has non-constant variance, called *hetroscedasticity*. The previous chapter, we have discussed penalization regression procedures under the assumption of *homoscedasticity*. In this chapter we will relax the assumption that the error expression variance is constant over observations and presume *hetroscedasticity* rather *homoscedasticity*, and using penalized weighted least squares rather than simple penalized least squares. We also compare the prediction and variables selection performance for Ridge, LASSO(Tibshirani, 1996), elastic net(Zou, and Hastie, 2005), Adaptive Lasso (Zou,2006) and suggested Lasso which have good oracle properties as discussed in previous chapter.

4.1 Introduction

In classical linear regression models it is assumed that the error part, often denoted by $\varepsilon_i^{'}$ are i.i distributed with mean 0 and constant variance σ^2 i.e., $E(\varepsilon_i^{'}/X_i)$=0, and $Var(\varepsilon_i^{'}/X_i) = \sigma^2$, where X_i means$X_{i1}, ..., X_{iP}$, for i=1,...,n.

Since, σ^2 is a measure of dispersion of the observed data value of the response variable (y is disperse about the regression line $\beta_1 X_1, ..., \beta_P X_P$)) homoscedasticity describes that the dispersion is similar across all data points. However, in various conditions, this supposition might be false. For instance, consider the sample of consumption expenditure of household and their income. Considering family with lower income do not have greater flexibility in spending, consumption designs among these low-income families may not differ very much. On the other side, rich households have a substantial deal of flexibility in spending. Some might be large-scale consumers; others might be wide-scale savers and invested a lot of their money in financial markets. This indicates that real consumption might be completely different from average expenditure. Thus, it is most probably that higher income families have a more dispersion about their average consumption than lower income households. Same condition is called *hetroscedasticity.* In the presence of *hetroscedasticity*, the OLS estimator is still unbiased, consistent and asymptotically normally distributed but its no more efficient.

One can identify the non-constant variances in the error terms, or *heteroscdasticity* by plotting the residuals against fitted values, if the plot shows some trend or a funnel shape then it is the indication of non-constant variances in the error terms, or *heteroscedasticity.* When we have *heteroscedasticity*, one possible solution is to transform the dependent variable Y such as $log(Y)$ or $\sqrt{Y}$, but one drawback is that it become concave function and as a result there will be problems in optimization in OLS or penalized least squares procedures. Therefore, we can use another method that is called generalized or weighted least square(WLS), applies the inverse of errors variances proportional to consider weights. We have used this idea in penalized regression procedures.

4.2 Consequences of Ignoring Non-constant Error Variance Problem

Consider a multiple linear regression model [17],

$$y_i = \beta_1 + \beta_2 X_{i1} +, ..., +\beta_p X_{ip} + \varepsilon_i \tag{4.1}$$

In above equation $Var(\varepsilon_i/Xi) = \sigma_i^2$ for i=1,...,n. So that, the error expression variances are disparate for all units of i.

Whenever one disregard *heteroscedasticity* and blindly applies OLS procedure to estimates β's the characteristics of unbiasedness and consistency still are not violated. However, OLS estimates in this problem are no more efficient. It is feasible to obtain an another unbiased linear estimate which has a low dispersion than estimates obtained from OLS method by utilize WLS. By simulation study, we will also examine the prediction accuracy and variables selection performance of different regularization and shrinkage methods in the presence of *heteroscedasticity*.

Inefficiency of OLS estimators:

$$y_i = \beta xi + \varepsilon_i \tag{4.2}$$

Where $Var[\varepsilon_i] = \sigma^2 w_i$, we consider that y_i and x_i are computed as deviations from their respective means, so that $E(y_i) = E(x_i) = 0$. Let suppose x_i represents the any predictor having n observations, and let Ω be a diagonal matrix whose diagonal elements are variances of error term. Now variance of the OLS estimator β is

$$Var[\tilde{\beta_{OLS}}] = \sigma^2(x^{'}x)^{-1}x^{'}\Omega x(x^{'}x)^{-1} = \frac{\sigma^2\sum_{i=1}^{n} x_i^2 w_i}{(\sum_{i=1}^{n} x_i^2)^2} \tag{4.3}$$

and Variance of WLS estimator is given by,

$$Var[\tilde{\beta_{WLS}}] = \sigma^2[x^{'}\Omega^{-1}x]^{-1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i^2/w_i)} \tag{4.4}$$

Clearly, $\frac{Var[\tilde{\beta_{OLS}}]}{Var[\tilde{\beta_{WLS}}]} > 1$, it displays that the achievement in efficiency obtain from WLS over OLS can be important. We can use this idea in penalized least squares.

Secondly, The standard error and covariances of the OLS estimates of the β's are become bias and inconsistent when heteroscedasticity is substantial but ignored. So that, the tests of hypotheses are not consider for further analysis due to wrong results. This inefficiency also badly affect forecasting, becuase the OLS estimates are no longer follow the well known property BLUE(also named, best linear unbiased estimator) and will be not applicable, therefor forecasts will also be not efficient. Now we summarizes a small simulation study that demonstrate the better fitting of WLS than OLS in heteroscedasticity. In this example,we generate 100 observations from the model,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4.5}$$

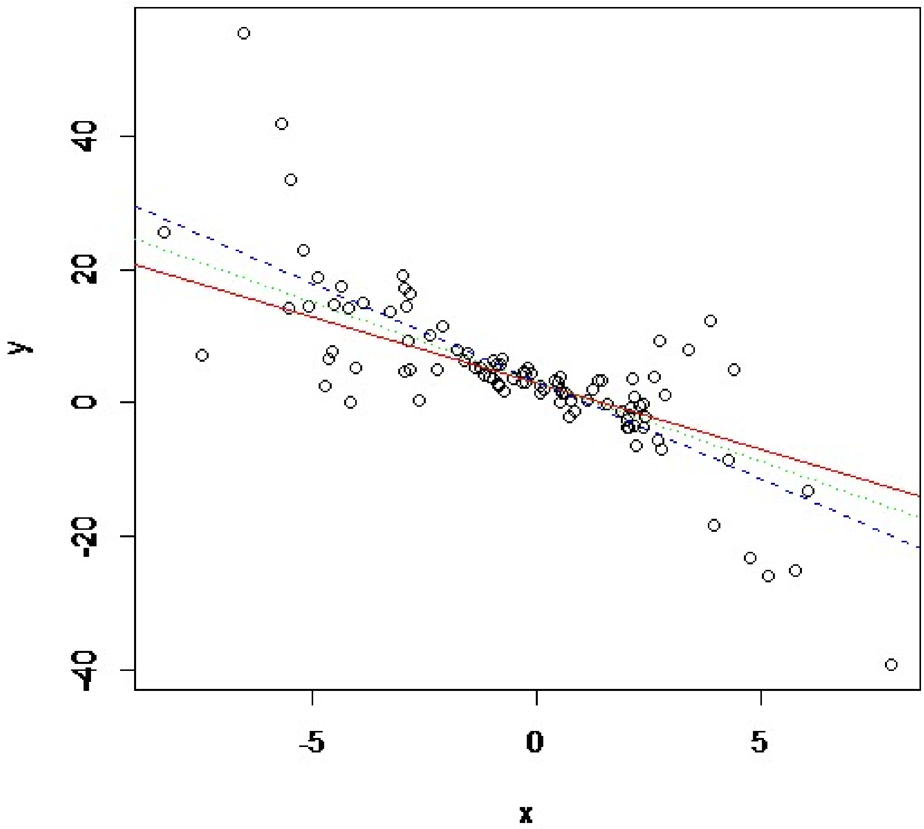


Figure 4.1: Graphical Comparison

Where $\beta_0 = 3$ and $\beta_1 = -2$, the predictor x from normal with mean 0 and variance 9. The error term ϵ_i generated from normal with mean 0 but its non-constant variance $(1 + 0.5x^2)$. In figure 4.1, the red fitted line shows the true regression line of equation(5). The blue curve indicates the OLS fit on simulated data set, and green line shows the WLS fitted line. From the figure 4.1 its clear that the WLS is approximate fit to true regression line and OLS is not good fit on this contaminated data set.

4.3 Penalized Weighted Least Squares and Variable Selection

Consider the linear regression model

$$y = X\beta + \varepsilon_i. \tag{4.6}$$

where y is an $n \times 1$ vector and X is an $n \times d$ matrix. As in the traditional linear regression model, we assume that ε_i are i.i.d normal with zero mean and constant variance (say, σ^2).

Lets relax the assumption that the error term has constant variance. Suppose ε_i is a random factor with mean $E(\varepsilon_i/X_i) = 0$ and $Var(\varepsilon_i/X_i) = \sigma_i^2$, for i=1,....,n. Thus,

$$E(\varepsilon_i \varepsilon_i^t) = \sigma^2 \Omega, \tag{4.7}$$

where Ω is a diagonal matrix. It will sometimes be useful to write,

$$\sigma_i^2 = \sigma^2 w_i \tag{4.8}$$

Now we consider the connection between the penalized weighted least squares and variables selection in linear econometric model. The WLS estimate is determining via minimizing,

$$L(\beta, w) = \sum_{i=1}^{n} w_i (y_i - X\beta)^2 \tag{4.9}$$

The OLS estimate is a special case of WLS when all weights becomes 1, i.e., $w_i = 1$. We can solve WLS by the same kind of algebra as we solve the OLS problem and obtain,

$$\tilde{\beta}^{WLS} = (X^T \Omega^{-1} X^T)^{-1} X^T \Omega^{-1} y \tag{4.10}$$

In present subsection we suppose that the columns of X matrix are orthonormal i.e, ($X^T X = 1$). Then penalized weighted least squares minimize,

$$W(\beta, w_i, \lambda) = ||w_i(y - X\beta)||^2 + \lambda \sum_{j=1}^{d} P_j(|\beta|) \tag{4.11}$$

If the penalty function $P_j(.)$ is equal to $\lambda \sum_{j=1}^{d} |\beta_j|$, then above equation equivalent to weighted LASSO problem.

Now solving by simple matrix algebra, we have

$$\tilde{\beta}^{WLASSO} = \tilde{\beta^{wols}} - \frac{\lambda}{2} sign(\tilde{\beta^{ols}}) \tag{4.12}$$

Clearly, weights are known and $w_i > 0$ so that signs are same for weighted and ordinary least squares estimates, hence the above equation can also be written as

$$\tilde{\beta}^{WLASSO} = \begin{cases} \tilde{\beta^{wols}} - \frac{\lambda}{2} & \text{if } \tilde{\beta^{ols}} \geq 0 \\ \tilde{\beta^{wols}} + \frac{\lambda}{2} & \text{if } \tilde{\beta^{ols}} < 0 \end{cases}$$

In condition, whenever the predictors matrix X is an orthogonal i.e, $X^T = X^{-1}$, then it simple to determine that for two model complexity control parameters (λ_1, λ_2) the naive elastic net provide solution easily for elastic net.

$$\tilde{\beta}(w - enet) = (1 + \lambda_2)\tilde{\beta}(w - naive) \tag{4.13}$$

The weighted adaptive LASSO is defined by,

$$\sum_{i=1}^{n} W^*(y_i - x_i\beta)^2 + \lambda_n \sum_{j=1}^{P} \tilde{w_i}|\beta| \tag{4.14}$$

Here W^* are heteroscedastic weights. Minimizing (14) with respect β leads to weighted penalized adaptive LASSO estimator of β.

4.4 Simulation Study

In this section we report some numerical experiments to assess the performance of modified method with Ridge regression, the LASSO, the elastic net and adaptive Lasso(H.Zou,2006), when error term has non-constant variances. In the present empirical study we examined different linear multiple models, $y = x^T\beta + \varepsilon_i$ here we assume that the error term has still normal but non-constant variances, or heteroscedasticity i.e., $(Var(\epsilon_i/X_i) = \sigma_i^2)$. In all considered problems, we calculated the weights for modified method utilizing Ridge regression estimates for optimum value of λ (λ is find by using 10-fold cross validation) rather than OLS coefficients. We show the numerical demonstration by following models.

Table 4.1: Simulation results of Model 1, Comparing the Median Prediction Mean Squarred Errors Based on 1000 Replications

		Avg.No. of 0 Coefficients	
Method	Med.PMSE	Correct	Incorrect
$w = 1/1 + 1.5x^2$			
Ridge	9.507 (0.117)	...	....
LASSO	8.961 (0.111)	2.724	0.161
Elastic Net	9.149 (0.114)	2.413	0.301
Adaptive LASSO	8.887 (0.112)	3.487	0.321
Proposed	8.465 (0.098)	3.963	0.155
w=1			
Ridge	11.86 (0.142)	...	....
LASSO	11.207 (0.14)	2.561	0.061
Elastic net	11.142 (0.146)	1.606	0.080
Adaptive LASSO	11.432 (0.157)	2.432	0.154
Proposed	11.22 (0.145)	2.067	0.210

Inside each model, our generated data set pertaining of a training and test data to compare the prediction accuracy and model selection performances of each regularization and shrinkage method. In each and every model, we simulated 1000 data sets and the numbers of zero coefficients selected correctly and incorrectly by each method are averaged for 100 best models using Bootstrapping with B=1000.

Model 1(large variances of residual). We let $y = x^T\beta + \varepsilon_i$, where the presupposed model parameters are $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$. The regressors $x_i(i = 1, 2, ..., n)$ are independently and identically normally distributed with zero mean and unit variance, and the pair wise correlation between them is set to be $corr(x_i, x_j) = 0.5^{|i-j|}$. Here we also assume that $\varepsilon_i \sim N(0, 1 + 1.5X^2)$ which tells us that the noise has a standard deviation that goes; $1 + 1.5X^2$.

Model 2 (small variances of noise). Model 2 is same as model 1 except that residuals have unequal but small variances, i.e, $\varepsilon_i \sim N(0, 1 + 0.5X^2)$ instead of $\varepsilon_i \sim N(0, 1 + 1.5X^2)$. Here, we have also compared the performances of each procedure in various sample sizes i.e, n=50,100,and 150.

Model 3 (Many small effects). Model 3 is same as Model 1 except that $\beta_j = 0.8$ for all j.

Table 4.1 summarizes the simulation results of model 1. Many important observations can be made from table 4.1. First we observed that the prediction performances of five variable selection and shrinkage methods in terms of weighted penalized criterion and without weighs in case of non-constant variances for residuals,or *heteroscedasticity*, weighted penalized regression approaches are better than using traditional approaches. First consider ridge regression, the median prediction error in case of weighted penalty is about 20% lower than that of median prediction error in case of without using weighted penalized regression, if error term has non-constant variances. Similarly, weighted LASSO has median prediction error is 20% lower than that of using simple LASSO, Elastic net has median prediction error is 18%, Adaptive LASSO (H.Zou, 2006) is 22% and proposed method is 25% lower than that of using simple methods respectively. Hence, maximum reduction in prediction error is occurred in proposed method.

Second, the proposed weigted penalized approach dominates all other methods in terms of prediction error(i.e, 8.465) when errors are heteroscedastic, as reflected in upper section of table 4.1.

Third, if we ignore the *heteroscadasticity* and use traditional penalized approaches we can see that any method can perform well any time, here elastic net performed better than alternative procedures concerning of their respective prediction accuracy.

Fourth, the standard errors (of the medians) estimated by using the bootstrap with B=500 re-samplings on the 1000 mean-squared error are shown in parentheses. It can be seen that the modified model determination method has smaller standard error, i.e., 0.098.

Now consider the variables selection performances of each method. In table 4.1 the column 3rd and 4th shows the numbers of zero coefficients estimated by each method correctly and incorrectly which we consider in model 1, that is $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$. There are five zero coefficients and three non-zero. The best method will that one which select about five correctly and 0 incorrectly coefficients. We can see that all methods performed very poor in terms of variables selection when the assumption of homoscedasticity is violated and still we used the standard penalized regression approaches. In case of using weights that are equal to inverse of variances of error term, clearly we can see that the proposed method perform very well, that is selected about 4 variables as zero, correctly, and almost one variable estimated as 0 by propose method incorrectly which is minimum ratio amongst others. Table 4.2, presents the implementation of the five methods for different

Table 4.2: Simulation results of Model 2, using different sample sizes

		Avg.No. of 0 Coefficients	
Method	Med.PMSE	Correct	Incorrect
n=150,P=8			
Ridge	3.037 (0.025)	....	....
LASSO	2.818 (0.024)	2.788	0
Elastic Net	2.862 (0.022)	2.250	0
Adaptive LASSO	2.808 (0.023)	3.800	0
Proposed	2.868 (0.023)	4.090	0
n=50,P=8			
Ridge	3.405 (0.06)	....	....
LASSO	3.106 (0.05)	2.471	0.021
Elastic net	3.202 (0.057)	2.073	0
Adaptive LASSO	2.936 (0.048)	3.424	0.017
Proposed	2.681 (0.042)	4.236	0

sample sizes. The median prediction mean squared errors(Med.PMSE) are a little smaller for large sample size(n=150) than median prediction mean squared errors for small sample size(n=50), also with same behavior of bootstrap standard errors. The variable selection performance are also almost same but for small sample size the LASSO and adaptive LASSO(H.Zou,2006) performed a little worse.

From table 4.3, it can be seen that when the error term is still non-constant variances but small than that of using in model 1, and the sample size remains the same i.e,. 100, proposed method performs well and it appreciably condense both model prediction error and model complexity. Elastic net and adaptive LASSO performs poorly, with regard of both test error and model selection.

One important finding is that when error term has unequal variances but small, then average number of zero coefficients estimated as 0 incorrectly is 0 by every procedure except adaptive LASSO.

In table 4.4, we have summarized the simulation outcomes of model 4.3 in which we consider many small non-zero coefficients. we find that the ridge regression performed better as compared to other methods. The elastic net has a little bit smaller test error than ridge regression but that selected incorrectly 2 variables as 0.

Table 4.3: Simulation results of Model 2, using $w = 1/1 + 0.5x^2$ are weights

		Avg.No. of 0 Coefficients	
Method	Med.PMSE	Correct	Incorrect
n=100,No weights			
Ridge	3.198(0.029)	....	....
LASSO	2.988(0.03)	2.60	0
Elastic Net	3.087(0.03)	1.97	0
Adaptive LASSO	3.007(0.032)	3.58	0.04
Proposed	2.935 (0.027)	4.48	0
n=100,W=w			
Ridge	3.052 (0.031)	...	....
LASSO	2.877 (0.029)	2.63	0
Elastic net	2.9 (0.028)	1.78	0
Adaptive LASSO	2.823 (0.028)	4.12	0
Proposed	2.771 (0.025)	4.23	0

Table 4.4: Simulation results of Model 3

Method	Med.PMSE	Avg.No. of Zero Coefficient
n=100,P=8		
Ridge	10.615 (0.135)	0
LASSO	10.733 (0.136)	2.34
Elastic Net	10.421 (0.139)	2
Adaptive LASSO	11.586 (0.153)	3.18
Proposed	11.052 (0.13)	2.7

Chapter 5

Assessing the Performance of Penalized Regression Methods in Generalized Linear Models

5.1 Introduction

In previous two chapters we have assumed that the error term ϵ_i is assumed to be normally distributed with mean, say μ, and variance σ^2. But in many instances, the error term ϵ_i is not normally distributed. Hence, generalized linear models(GLM's) are used instead of classical linear models. GLM's were first proposed by Nelder and Wedderburn (1972). A further development is given by McCullagh and Nelder (1989). GLM's are most commonly used to binary, or count data so our center of interest will be on models for these types of data. According to Fahrmeir and Tutz, a GLM has two features:

1.The distributional assumption forms that, y_i given X_i is conditionally independent, i.e, $f(y_i|X_i) = f(X_i)f(y_i)$.

2. The above conditional distribution of y_i belongs to a simple exponential family with density function can be written as,

$$f(y_i|\theta_i, \phi) = exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\}. \tag{5.1}$$

In above density θ_i and ϕ are parameters, ϕ is an additional scale or dispersion parameter and $b(.)$, $c(.)$ are known functions of the exponential class. We also assumed that the expectation $E(y_i|X_i) = \mu_i$ and the parameter ϕ is independent of the ith observation.

Let (y,X) denote a set of a binary response variable y and X be a p-dimensional predictor vector, the distribution of each y_i is (5.1). Then the relationship between the mean of the dependent variable y, denoted as μ_i and the linear function X $\psi_i = X_i^T \beta$ is find by the structural supposition. Therefore, [27]

$$\mu_i = h(\psi_i), \psi_i = g(\mu_i) = h^{-1}(\mu_i) \tag{5.2}$$

In last equation $h(.)$ is the outcome function and $g(.)$, the reciprocal of outcome function h, is well known quantity in generalized linear econometric models named link function, for instance, in binomial density this, $g(.) = log(.)$. This link function has more importance in GLM models because without using this the following issues occur [27].

1. The range of the parameter of y variable distribution may not be same with the range of predicted parameter. For instance, p range for a binomial distribution is between 0 and 1–i.e, logistic regression, and in predicted it may be negative infinity to positive infinity.
2. The relationship between y and X not to be the simple linear form.
3. The third substantial cause is that without choosing a proper link function, the variances of the error term will not be constant.

Therefore, a one-to-one continuous differentiable link function will be establish. The natural parameter θ_i is a function of the expected value of μ_i, i.e. $\theta_i = \theta(\mu_i)$. Moreover, the mean of the response factor y is become $\mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}$ and variance $\upsilon(\mu_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}$. The variance function $\upsilon(\mu_i)$ results from $var(y_i|X_i) = \psi \upsilon(\mu_i)$, where $\upsilon(\mu_i)$ is the variance of the response variable y_i. The identification of the mean $\mu_i = h(\psi_i)$ implies a definite variance form as these two are depend on derivatives of $b(\theta_i)$. If θ is directly associated to the predictor vector X, the connection function is called natural or canonical link form and set by [27]

$$\theta(\mu_i) = \psi_i, g(\mu_i) = \theta(\mu_i) \tag{5.3}$$

The binomial and Poisson distributions are also representatives of the exponential class. The before mentioned distributions have the below natural link functions:

$\psi_i = log(\frac{\mu_i}{1-\mu_i})$ for the Bernoulli distribution.

$\psi_i = log\mu_i$ for the Poisson distribution.

Besides these distributions we can also use the Gamma and multinomial distribution for generalized linear models. We will use these link functions to generate response variable values in simulation study.

5.1.1 Maximum Liklihood Estimation

Maximum likelihood is the estimation process for solving the generalized linear models. Whereas the dependent variables follow to an exponential class (1), and we take log of the likelihood form of observations $y_1, ..., y_n$ so we get [27],

$$l(\theta) = \sum_{i=1}^{n} l_i(\theta_i) = \sum_{i=1}^{n} logf(y_i|\theta_i, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} \tag{5.4}$$

Where, the function, $c(y_i, \phi)$ is neglected because it does not effect by θ_i and free of θ_i. The first derivative of the log-likelihood form as,

$$m(\theta) = \frac{\partial l(\theta)}{\partial \theta} \tag{5.5}$$

This function is consists of more than one unknown parameters, i.e. θ, μ, and ψ, therefore we will be using Fisher Information Matrix approach to solve equation (5). The matrix inscription is helpful for the calculation of the maximum likelihood problem. This types of estimates are derived by solving the question, $m(\hat{\theta}) = 0$.

In the context of logistic regression it is essential to mention the terminology of odds and odds ratios as parallel to probabilities and probability ratios. Suppose E be an event has probability π and so the complementary event E^c has probability $1 - \pi$. Then the odds are defined as the ratio of these two probabilities, i.e.

$$odds(E) = \frac{P(E)}{P(E^c)} \tag{5.6}$$

It is well known that the probabilities are restricted to the interval [0,1], while odds lies in the interval [0,∞]. By taking logarithms of last equation we get the log odds, which are commonly enable to take the values $-\infty$ and $+\infty$ in order to construct an one-to-one relationship with the

Table 5.1: 2 × 2 contingency table

	Y=0	Y=1
Z=0	1363	1625
Z=1	1475	697

respective probabilities 0 and 1. Note that in case of logistic probabilities the odds are,

$$odds = \frac{exp(\psi_i)/1 + exp(\psi_i)}{1 - [exp(\psi_i)/1 + exp(\psi_i)]} = exp(\psi_i) \tag{5.7}$$

Thus, for the log odds we get,

$$log\{odds(\psi_i)\} = log.exp(\psi_i) = \psi_i. \tag{5.8}$$

Finally, the odds ratio, as the name propose, is simply the ratio of the odds of two events and may be considered as a measure of the effect size reporting the degree of association between two events. For the sake of intelligibility and a better understanding of the basic idea behind odds ratios we shall consider the following example.

Assuming a logit model with only one binary covariate, we may define Z=1 for, wearing a ski helmet, versus Z=0, if the skier does not wear a helmet. The response variable Y equals 1, if a skier supported severe head injuries from a ski accident and 0, if the head injuries were not severe, thus Y incorporate only accidents with head injuries.

The underscore data, taken from the $Freizeitunfallstatistik$ 2007, Kuratorium fur Schutz and Sicherhiet, is presented in table 5.1.

However, keep in mind that the number of accidents with helmets is an approximate, because unfortunately the exact number was not recorded in statistic. Nevertheless, we may calculate the odds like the following:

$Odds(Y|Z = 0) = 1625/1363 = 1.19$ and $odds(Y|Z = 1) = 697/1475 = 0.47$, respectively, and the odds ratio is $odds(Y|Z = 0)/odds(Y|Z = 1) = 1.19/0.47 = 2.53$. The result may be explicated that the risk of severe head injury not wearing a helmet is 2.53 times as more as for other skiers wearing helmets.

5.2 Models for Binary Response

For modeling the binary dependent variable the largest general applied machine learning tools are Logistic regression models. The prevailing models have simple computational processes in constructing regression models and more interpretable. Because of high interpretability and therefore may be applied in situations where explanation of the model is vital than prediction accuracy. Logistic regression models was developed by David Cox in 1958.

Let $y_i \in \{0, 1\}$ be a binary outcome, then it could be modeled at using the Bernoulli distribution, i.e. $y_i \sim Ber(1, \pi_i)$. So that, the dependent variable probability is determined through [1]

$$E(y_i|x_i) = P(y_i = 1|x_i) = \mu_i = \pi_i \tag{5.9}$$

and $\theta(\pi_i) = log(\frac{\pi_i}{1-\pi_i})$, $b(\theta_i) = log(1+exp(\theta_i)) = log(1-\pi_i)$ and ϕ is equal to 1, are the components of the exponential class. By choosing such link function, in term of log, leads to mapping from (0,1)→ $(-\infty, +\infty)$. This also implies the variance structure $var(y_i|x_i) = \pi_i(1 - \pi_i)$. The following model and their proportional link functions are very simple to connect with the probability of categorical dependent variable π_i to the linear function $\psi_i = x_i^T \beta$.

The logit or logistic model is formed through canonical link function [27],

$$g(\pi_i) = log(\frac{\pi_i}{1 - \pi_i}) = \psi_i \tag{5.10}$$

So that, given the derived outcome variable function (the conditional probability) we obtain the logistic or logit distribution form as [27],

$$\pi_i = h(\psi_i) = \frac{exp(\psi_i)}{1 + exp(\psi_i)} \tag{5.11}$$

where $\psi_i = x_i^T \beta$ and one observe P explanatory variables (or covariates), $x \in R^p$ and a binary regressand (or output) variable $y \in \{0, 1\}$ and therefor express the group membership of the recognized vector of regressors. Here $\beta \in R^p$ are their respective values of coefficients. A positive value of the estimated coefficient for a covariate indicates that this covariate or independent variable is related with a high probability of the predicted $(y = 1)$, on the other hand, a negative

[1] If P(y=1/x)=π_i, then P(y=0/x)=1-π_i

estimated coefficients minimizes the probability of strong relationship between outcome variable and regressors. A regressor with estimated coefficient of 0 has no impact on the probability of dependence and must preferably be eliminated from the consider regression model.

5.3 Constrained Logistic Regression

Usually, we know that maximum likelihood procedure mostly overestimates logit regression parameters leads to models that estimate badly. With regard to face this difficulty several procedures that condense or shrunk the coefficients and work automatic predictor selection have been offered. In current thesis we discusses the performance of modified shrinkage and regressor selection approach against already recommended penalized logit regression models which have great applications in nowadays, that are: (1) Ridge regression, (2) the Least Absolute Shrinkage and Selector Operator(LASSO), (3) the elastic net, and (4) the Adaptive Lasso and so on.

5.3.1 Ridge Regression

This approach, maximizes the log function of likelihood subject to a constraint on the magnitude of the regression parameters. This consequences to continuously diminishing of the maximum likelihood parameter estimates to 0 and doing this leads to enhance prediction performance. In this context the stated procedure form as [11, pp.1348-1360],

$$\hat{\beta}^{ridge} = argmax_{\beta}\{l(\beta) - \lambda\sum_{j=1}^{p}\beta_j^2\} \tag{5.12}$$

Generally the last problem is maximized, but we can also written in form given below of minimizing,

$$\hat{\beta}^{ridge} = argmin_{\beta}\{-l(\beta) + \lambda\sum_{j=1}^{p}\beta_j^2\} \tag{5.13}$$

with respect to β. To received a penalized maximum likelihood value for β, we minimize (11) with regards to β for specified thresholding parameter λ. Ridge model can mostly solved models that predict good but cannot produce sparse or simple models, a sparse model actually that one in which greater number of coefficients are estimated as zero. Ridge model either condense all estimates of parameters toward 0, and hence keeps all in the resulting model.

5.3.2 Least Absolute Shrinkage and Selection Operator

This method of shrinkage and predictor preferences, that is LASSO, executes a L_1 constraint penalty on the econometric model parameters in place of L_2 constraint regression. This inspires a methodology that simultaneously works coefficient shrinkage and suitably predictor selection accordingly. In GLM's model, LASSO estimator is form [11, pp.1348-1360],

$$\hat{\beta}^{LASSO} = argmax_{\beta}\{l(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|\} \tag{5.14}$$

It is equivalently to minimizing,

$$\hat{\beta}^{LASSO} = argmin_{\beta}\{-l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|\} \tag{5.15}$$

In last form $\lambda \geq 0$ regulates the volume of condensing or shrinkage. The major edge or benefit of LASSO over ridge model is that it can produce sparse models, these types of representations have simple in explanation. Moreover, the LASSO may displays unpleasant performance in the existence of extreme correlations among explanatory variables. For instance, whenever there is a class of regressors and that are strong relationships among them, the LASSO works to randomly incorporate one covariate from that class or category and neglected the other regressors. Furthermore, simulation experiments indicates that ridge procedure outperforms the LASSO in linear econometric models whenever there are small(near to 0) multiple explanatory variables among which strong correlation.

5.3.3 Elastic Net

In 2005, Zou and Hastie have established a fresh penalized regression formula, named as above, to address the bad capability of LASSO particularly in case of several associated regressors. Such as the LASSO, this procedure may generate sparse models and also reduce estimated coefficients. Moreover, when we have a category of strongly correlated independent variable, the elastic net maintain the whole category of all predictors because these all are significant, and also enhance prediction accuracy. This constraint regression procedure is become due to the collaboration of the

LASSO and ridge regression models, designate as

$$\hat{\beta}^{Enet} = argmax_{\beta}\{l(\beta) - \lambda_1 \sum_{j=1}^{p} \beta_j^2 - \lambda_2 \sum_{j=1}^{p} |\beta_j|\} \tag{5.16}$$

Or

$$\hat{\beta}^{Enet} = argmin_{\beta}\{-l(\beta) + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|\} \tag{5.17}$$

Here the two unknowns $(\lambda_1, \lambda_2) \geq 0$ are model complexity control parameters. Besides ridge and LASSO estimation methods, elastic net needs inference of both constraint parameters, make this approach a little laborious. Furthermore, it has been noticed that the consider method may over condense the regression estimates against its true value in linear models. But this problem has been solved latterly by them, to convert navie elastic net to modified.

5.3.4 Adaptive LASSO

Hui Zou, was proposed the adaptive Lasso(2006) estimator. The LASSO(1996) is a useful and fascinating procedure for parameter estimation and in feature selection. The prevailing method has not consistent in predictor selection even for large sample size. But sometimes it become consistent under specific restrictive situations. Nevertheless adaptive Lasso is a new version of existing method, in this formula OLS estimates consider as weights and are utilized in various constraint coefficients algorithm in the L_1 penalty. Then adaptive lasso appreciates the oracle properties; namely, it works really good whenever the true consider model were given in advance. The penalized likelihood adaptive Lasso estimator is

$$\hat{\beta}^{AL} = argmax_{\beta}\{l(\beta) - \lambda_n \sum_{j=1}^{p} \hat{w_j}|\beta_j|\} \tag{5.18}$$

Or equivalently,

$$\hat{\beta}^{AL} = argmin_{\beta}\{-l(\beta) + \lambda_n \sum_{j=1}^{p} \hat{w_j}|\beta_j|\} \tag{5.19}$$

Where $\hat{w} = \frac{1}{|\beta^*|^{\gamma}}$ are define as weights, H.Zou, use OLS estimators as weights, i.e. $\beta^*(OLS)$ is a consistent estimator for β. Here (λ_n, γ) are tuning parameters, and we wish to obtain a best pair of (λ_n, γ) by using two-dimensional cross-validation. When these weights are find out efficiently from given data set then it have follows the oracle properties.

5.3.5 Proposed Approach

In high-dimensional data environment(e.g., microarrays data sets), often denoted by a symbol $P \gg n$; then it is not possible to obtain the OLS estimator from underline data set and utilizing these as weights in adaptive Lasso(2006). So that, we have modify its formula and suggest a new version of the adaptive LASSO. In place of OLS estimates we use the ridge regression coefficients which are consistent under some conditions(i.e, in augmented data). We also observe that one extra tuning parameter the L_2 regularization parameter is incorporate in this new type of feature selection method. A shape of the penalized likelihood function for this approach mus equivalently be written like,

$$\hat{\beta}^{PAL} = argmax_{\beta}\{l(\beta) - \lambda_n \sum_{j=1}^{p} \hat{w_j^*}|\beta_j|\} \tag{5.20}$$

Where the data-dependent weights are, i.e, $\hat{w_j^*} = \frac{1}{|\hat{\beta}^{ridge}|}$.

5.3.6 Our Solution: L_1 Proposed Penalty

We consider a binary response variable y(i.e, logistic regression), then its density function will be a exponential family of binomial distribution $Bin(n, \beta)$. In above penalized function $l(\beta)$ represents the likelihood of binomial density. Solving the likelihood function and rearranging the same terms we get,

$$l(\beta_j) = ylog(\frac{\beta_j}{1-\beta_j}) + nlog(1-\beta_j) + C \tag{5.21}$$

Where C is constant because it free of β. Now, the complete penalty function of the modified adaptive LASSO become,

$$L(\beta, \lambda) = l(\beta_j) + P(\cdot) \tag{5.22}$$

where $P(\cdot)$ is the modified penalty depends on λ and β's. The optimization of last equation to get the maximum likelihood estimator of β through differentiation is difficult, because the log-likelihood function $l(\beta_j)$ is a concave function. Though the last equation is look like usual penalized least squares counterparts, but non-linear in β therefore, iterative methods are needed to solve it. For example, applying the $Newton-Raphson\ algorithm$, approximates the target function by a second order Taylor series, and optimizes that approximation. A NewtonRaphson step is repeated until convergence at the optimum, resulting in an iteratively reweighed least squares (IRLS) algorithm.

5.3.7 The IRLS Algorithm

IRLS abbreviated is iteratively re-weighted least squares. The Newton-Raphson procedure is the common regime for maximization of maximum likelihood in GLM,s. He developed IRLS algorithm for this purpose. In IRLS method the 2^{nd} order taylor approximation is used, and he needs the 2^{nd} derivative of function $l(\beta)$. After performing this process we construct the matrix called hessian, denoted by H, from these derivatives. In this subsection we consider the H matrix of the proposed penalized log likelihood will be actually negative, and used an IRLS algorithm for solving to find the estimators of proposed method. The general form of Newton-Raphson is simple and worked iteratively to finds coefficients such as,

$$y_{n+1} = y_n - \frac{f(y_n)}{[\frac{\partial f(y_n)}{\partial \beta}]} \tag{5.23}$$

Using this idea we have

$$\beta_{new} = \beta_{old} + (-\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T})^{-1} \frac{\partial l(\beta)}{\partial \beta} \tag{5.24}$$

Here $\frac{\partial l(\beta)}{\partial \beta} = X^T(Y-P) - \lambda \hat{w_j^*} sign|\beta|$ and $\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = H = -X^T DX$. Where D is diagonal matrix. The major problem here is that the H matrix is not in full rank, due to this issue the optimization problem become difficult. Therefore, we shall consider the gradient ascent algorithm, which includes all the steps of Newton-Raphson method. In this method we used the gradient, which is discussed in appendix, then we obtain the H matrix in full rank. The detail is given in appendix.

5.4 A Simulation Study

In this section we shall generated data sets via simulation process under four distinct scenarios basically considered by Zou, and H.Hastie(2005) and replicated here for expediency. In all four scenarios, we shall randomly break the generated data set into training and test parts. All regression models and thresholding values, λ's, are evaluated through only training data set. The prediction capability of all considered procedures are calculated entirely on the test data. Each approach is replicated for 1000 iterations.

The four scenarios are:

I. Example 1: In this example, we simulated the data sets comprising of 100 observations from the model $Y \sim Bernoulli\{p(x^T\beta)\}$, whereas $p(\psi) = \frac{exp(\psi)}{1+exp(\psi)}$. The true regression coefficients were set to $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, and the components of x are generated from standard gaussian. The pairwise correlation between predictors i and j is $\rho^{|i-j|}$ for all i and j, with $\rho = 0.5$. Model evaluation error are determined via 1000 Monte Carlo simulations.

II. Example 2: Just like Example 1, besides that $\beta_i = 0.85$ for all i.

III. Example 3: The n=400 observations are generated from the model given an example 1, and $\rho = 0.5$ for all i and j. The true coefficients (P=40) are assumed,

$\beta = \{(0, ..., 0)_{10}, (2, ..., 2)_{10}, (0, ..., 0)_{10}, (2, ..., 2)_{10}\}$

IV. Example 4: In this case we have generated data sets with grouped explanatory variables. The n=400 observations are simulated. The values of considered coefficients(P=40) are set such as,

$\beta = \{(3, ..., 3)_{10}, (2, ..., 2)_{5}, (0, ..., 0)_{25}\}$

The predictors are generated such as:

$x_i = Z_1 + \epsilon_i$, where $Z_1 \sim N(0, 1)$, for i=1,....,5

$x_i = Z_2 + \epsilon_i$, where $Z_2 \sim N(0, 1)$, for i=6,....,10

$x_i = Z_3 + \epsilon_i$, where $Z_3 \sim N(0, 1)$, for i=11,....,15

and the remaining 25 regressors are generated as, $x_i \sim N(0, 1)$ for i=16,....,40; where N(0,1) represent the standard gaussian distribution and $\epsilon_i \sim N(0, 0.01)$ for i=1,...,15.

The median prediction mean square errors(MPMSE) for the four simulation scenarios is shown in Table 5.2. In scenario (1), the adaptive LASSO and proposed method have very poor performance with regards to prediction accuracy and also have the largest bootstrap standard errors. In contrast, LASSO is performed better than all other penalized logistic regression procedures, i.e, have minimum median test error and bootstrap standard error too.

Table 5.2: Median of test errors for the simulated examples and six methods based on 1000 replications; bootstrap estimates of standard errors given in parentheses

Method	Simulation			
	Example1	Example2	Example3	Example4
Ridge	1.605(0.027)	1.264(0.025)	19.405(0.214)	0.156(0.001)
LASSO	1.323(0.02)	1.205(0.017)	2.686(0.015)	0.136(0.002)
Elastic Net	1.755(0.024)	1.992(0.026)	5.638(0.031)	0.123(0.002)
Adaptive LASSO	5.896(2.786)	4.137(0.785)	....	....
Propose	5.583(0.071)	1.233(0.019)	2.724(0.016)	0.104(0.001)
Oracle	0.309(0.013)	1.193(0.028)	1.036(0.02)	0.078(0.001)

In example 2, we have considered many small non-zero effects. The three methods that are LASSO, ridge and proposed procedure, performed relatively well in term of prediction error and perform just like oracle estimator(i.e, 1.193). In example 3, clearly we can see that ridge regression is the largest median prediction error and contrast, the LASSO and proposed method perform very well. It is interesting to see that the adaptive LASSO estimator cannot possible to calculate in example 3 and example 4.

In grouped explanatory variables situation presented in Example 4, the modified method outperforms the all other approaches, and can be seen that the performance of proposed procedure is expected to be good as that of the oracle estimator. The test failure rate of the suggested method is around 15% lower than that of elastic net, 24% and 33% lower than LASSO and ridge regression respectively. The bootstrap standard errors of the ridge regression, the propose method and oracle estimator are exactly same, i.e, 0.001.

The variables selection performance of all procedures are demonstrated in table 5.3 and table 5.4. The average number of correctly estimated coefficients as 0 is reported in table 5.3, and incorrectly estimated 0, when in given model it is not equal to 0, in table 5.4 over the 100 best models. The ridge regression results are omitted because it shrunk the coefficients but not exactly 0. From table 4.3, it can be seen that in example 1 the proposed method perform best and its performance is almost same with oracle method, i.e, its selected 4.45 average number of 0 coefficients correctly while oracle is 5. The LASSO, elastic net and adaptive LASSO presents the mean number of correctly estimated as 0 coefficients are 3.42,3 and 3.604 respectively. Now from table 5.4, the

Table 5.3: Mean Number of Correctly zero Estimated Coefficients

	Simulation	Results		
Method	Example1	Example2	Example3	Example4
LASSO	3.42	0	20	20.381
Elastic Net	3	0	9.276	18.2
Adaptive LASSO	3.604	0	—	—
Propose LASSO	4.4	0	13.813	10.1
Oracle	5	0	20	25

proposed method outperform the all other methods in example 1 and performs exactly same as oracle. The adaptive LASSO gives very poor results as compared to other penalized logistic regression. Hence, proposed procedure performs the best and it significantly reduces both, the model error and model complexity.

In example 2 we consider many non-zero small effects, from table 5.3 it can be seen that all the models perform well and no one coefficients estimated exactly zero, but from table 5.4, we can see that all the penalized procedures set some average number of coefficients erroneously to 0, except ridge regression, and still proposed method gives satisfactory results after ridge regression. Hence, in example 2 ridge regression perform best than other penalized logistic regression procedure.

In example 4.3, we set 20 coefficients exactly 0 and 20 coefficient non-zero, i.e, 2. The column 4 of Table 5.3 and 5.4 presents the performance of the four methods in variable selection of scenario 3. LASSO can correctly identify the 20 significant variables just like oracle(see table 5.3), but perform very worse when it estimate some coefficients exactly 0 in real that not set to 0 in given model. All other method perform just as oracle in table 5.4. The suggested method perform well in example 3, as we can see from table 5.3 that it selected average number of about 14 variables correctly to 0. LASSO performed poorly in Example 4 when compared to propose and the elastic net penalized approaches. As we can see that the proposed approach achieved the lowest average test error here(see in table 5.2), but in variable selection it is a little dominated by elastic net on grouped predictor variables in Example 4.

Table 5.4: Average Number of Inorrectly zero Estimated Coefficients

	Simulation	Results		
Method	Example1	Example2	Example3	Example4
LASSO	0.06	1.42	10.45	9.523
Elastic Net	0.1	1.636	0	0
Adaptive LASSO	0.509	3.34	—	—
Propose LASSO	0	1.8	0	0
Oracle	0	0	0	0

5.5 Regression Models for Count Data

A well known distribution for count data is the Poisson probability density function with parameter $\alpha_i > 0$ and output variable is express as $y_i \in \{0, 1, 2, ...\}$. There are many real life situations in which the response(dependent variable) is of the count shape, for instance, the number of holidays received by a Director from his department in per month, the number of catches by a sportsman in a test match, the number of cancer patients diagnostic per year, the number of accidents on motorway per week, the counting of students admitted in a school per year, the number of months spent in a foreign country in a given period, the number of deaths due to suicides attacks in a country per year, and so on. The underline variable in every example mentioned before is clearly discrete in nature, assuming entirely a finite number of units. Occasionally count data may also describe is rare, or infrequent, happenings, that is to say, taking 6 or more wickets by a bowler in a one day cricket match, the number of patients whose O negative blood group observed in a hospital per day, or, having one or more children without able to walking in a village observed per year.
We modeling the said circumstances exactly as the Bernoulli distribution was selected to model the yes/no response in the logistic regression approach, the probability density function that is particularly adjusted for count data is the Poisson probability distribution.
The parameters of the exponential family are specified as $\psi(\alpha_i) = log(\alpha_i)$, $b(\psi_i) = exp(\psi_i)$ and $\phi = 1$. Because of the 0 or positive values of the dependent variable, a commonly used link function is different from the link function implemented in the logistic model. Remaining penalized algorithms are same for the poisson distribution that were used in logistic regression, except the

Table 5.5: Simulation results Based on 1000 Replications; inside parentheses the numbers indicates standard error estimates by Bootstrapping with B=500 repeated re sampling.

		Avg.No. of 0 Coefficients	
Method	Med.PMSE	Correct	Incorrect
Ridge	3.078(0.036)	...	...
LASSO	3.135(0.035)	3.24	0.7
Elastic Net	3.192(0.037)	2.7	0.5
Adaptive LASSO	3.351 (0.037)	2.62	0.27
Proposed	3.289(0.031)	4	0.27
Oracle	1.090(0.015)	5	0

log-likelihood $l(\beta)$, for poisson probability distribution.

5.5.1 A Simulation Experiment

In current subsection we present a numerical study to differentiate the modified method with adaptive Lasso(H.Zou, 2006), LASSO, elastic net and ridge regression in Poisson regression. This numerical experiments are found on the Poisson model with the log-link is utilize as specific function, i.e, $y_i \sim Poisson(\alpha_i)$ with $\alpha_i = exp(x_i^T \beta)$.

In this experiment we consider sample size n=100 and predictor dimension p=8. We let $\beta = (0.45, 0.25, 0, 0, 0.3, 0, 0, 0)$. The predictors x_j for j=1,...,p are generated i.i.d gaussian with mean zero and unit variance. The association in pairing between covariates x_j and x_k is to form $corr(x_j, x_k) = 0.5^{|j-k|}$.

Table 5.5 summarize the prediction and variable selection performance. The findings shows that in this experiment the ridge regression has minimum prediction error i.e, 3.078, but it has selected all the variables. LASSO perform good after ridge in term of prediction accuracy but it has also poor performance in variables selection than proposed method. The findings shows that the propose method has 3.289 test error, and average numbers of 4 variables selected correctly, 0.27 incorrectly which is very near to oracle estimator. It can also be seen that the propose method is minimum bootstrap standard error(in parentheses) among all other procedures.

Table 5.6: Predictor coefficients estimates and classification test error solutions, for various shrinkage and variable selection procedures implemented to the stock market data. The dashed records corresponds to covariates not selected.

Variable	Ridge	LASSO	E.net	A.LASSO	Proposed
Intercept	-41.440	-35.993	-91.372	0.667	0.115
Year	0.021	0.018	0.046	....	....
Lag1	-0.084	-0.211	-0.280	....	....
Lag2	-0.096	-0.040	-0.198	....	....
Lag3	0.065	....	-0.033	....	....
Lag4	0.009	0.046	0.109	....	....
Lag5	0.099	0.146	0.236	0.123	....
Volume	0.055	0.196	0.329	....	....
Today	1.711	15.596	12.667	33.948	8.975
Test Error	0.5063	0.5040	0.5074	0.5017	0.5000

5.6 Real Data Examples

5.6.1 The Stock Market Data

In present segment we examined the achievements of all five procedures on the stock market data. This data set comprises of percentage returns for the S&P 500 stock index on to 1,250 days, from the starting point of 2001 until the end of 2005. For everyone date, we have recorded the percentage earns for each of the 5 previous marketing days, Lag1 through Lag5. We have also registered Volume (the number of shares traded on the previous day, in billions), Today (the percentage earnings on the date in question) and Direction (whether the market was Up or Down on this date).

The data set contains of 1250 observations and 9 variables. The binary outcome variable y (i.e, Direction) is 1 for which the market is up, and 0 otherwise. The other 8 variables are consider as predictors. During analysis, the data file is randomly separated into training, and testing sets. The train set consists of 375 observations and test data contains the remaining 875 observations. Table 6 presents the classification error on test data, estimation and variable selection performance. The results are shows that the test error rate of proposed method is minimum than other methods, i.e, 50% of the daily movements have been correctly predicted. Now consider the variable selection performance, we can identify that, ridge approach and elastic net selected all variables, LASSO select one variable(i.e, Lag3) is 0, adaptive LASSO estimated six regression coefficients are 0 and proposed method set seven coefficient is 0, which produce most sparse model.

5.7 Microarray Classification and Gene Selection

A prime implementation of DNA micro-array data is cancer category classification. Cancer is a terminology that describes to uncontrolled cellular division, growth and outspread of abnormal cells. It can transpire in complete human body parts. As stated by the world health organization(W.H.O), cancer is a disease that warnings human lives and causes the second highest rate of death globally. In cancer treatment or therapy, the classification or categorization of normal and abnormal structures of the cells be one of the very crucial and influential operations in the time of diagnosis of cancer. Currently, the application of proficient classifier methodologies in cancer diagnosis is raising. The principal objective of these classifiers, such as support vector machine(S.V.M) and some others, is to withdraw the beneficial learning outcomes from previous diagnosis evidences[41].

So the key problem of micro-array data is high-dimensionality, that is the number of genes(predictors), P, surpasses the number of tissues or patients(n). Over fitting, prediction accuracy and multi-collinearity are highly important and common issues in high magnitude data sets whenever we implemented statistical learning procedures for categorization or classification. Stated problems formulates the different micro-array classification techniques much difficult.

From the biological perspectives, among thousands of genes hardly a small subgroup of genes is greatly responsible for a focus disease, and a lot of genes are insignificant or irrelevant and not consider to cancer classification. The non significant genes may cause error and decrease the classification accuracy. Furthermore, from the statistics point of view, large numbers of genes may introduce over fitting and may badly effect the classification capabilities. Many statistical techniques have been successfully applied in the area of cancer classification in micro-array data set. Among them, penalized logistic regression methods are considered as useful and widely used in high-dimensional cancer classification. These approaches involve different types of penalties. One

Table 5.7: Summary of the leukaemia classification results

Method	10-fold CV error	Test error	Number of genes selected
Ridge	0/38	8/34	All
LASSO	1/38	2/34	2
Elastic Net	2/38	3/34	39
Propose LASSO	1/38	0/34	51

of them is the L_1 penalty, but it has three drawbacks described by Zou and Hastie(2005). They suggested the elastic net penalty, but it does not have the oracle properties.

Due to finally mentioned obstacle, the oracle properties, H.Zou introduced the adaptive Lasso(2006) method in its formula OLS estimates are utilized the adaptive weights. But in high-dimensional classification data it faces some practical problems in choosing weights. Therefore, we use the ridge coefficients just a primary estimator in resolve the adaptive Lasso[43].

5.7.1 Leukaemia Data

The micro-array leukaemia data comprise of 7129 genes(predictors) and 72 samples(n) of patients. In the training data part, there are 38 samples, among which 27 are type 1 leukaemia i.e, acute lymphoblastic leukaemia(ALL) and 11 are type 2 leukaemia named as, acute myeloid leukaemia(AML). In the test data set, there are 34 samples, in which 20 are type 1 leukaemia(ALL) and 14 are type 2 leukaemia(AML) [43]. The basic objective is to make a diagnostic principle stand on the expression level of those 7219 genes to predict the type of leukaemia. We first want to coded the binary response variable i.e, the type of leukaemia(ALL/AML), as a 0-1. We coded the ALL type of leukaemia as 0, and AML 1. We have used the indicator function $I(.)$ i.e, $I(fitted - value > 0.5)$, for classification. We have applied ten fold cross validation to determine the complexity control parameters.

Table 5.7 presents the classification performance in the *holdwith* and *holdout* data sets, and number of important genes selected. The results are shown that the proposed method has good classification accuracy on test data and gives 0 classification error as compared to LASSO and elastic net. The ridge regression provided worse test classification accuracy. The classification performance on training data set of all four methods are almost same, but ridge regression do well. The variable selection performance shows that LASSO select 2 variables and produce extremely sparse model, which is very bad work to select only two predictors from thousands. The elastic net selected 39 and proposed method 51 genes. Procedures choosing many predictors turn to over-fit the information set. Therefore, procedures with a small-scale number of nominated predictors are desired, but it does not means that the simplest model, which causes high biasedness as we can see from table 6 the LASSO result.

Chapter 6

Summary, Conclusion and Recommendations

A conclusion is the place where you get tired of thinking [1].

In this last chapter we summarize and conclude our findings. We shall also mention recommendations in the last section of this chapter.

6.1 Summary of Conclusion

In the context of regularized analysis, there are two goals: model selection and assessment of the selected model. Our research work is related at comparing the performance of proposed penalized method with LASSO, Ridge regression, elastic net and adaptive LASSO through simulation results on separate regression models. We also have applied these penalized approaches to the real data sets and compared the respective performances. The simulation results in chapter 2 demonstrate that the proposed method work well on low dimensional data set by taking into account the prediction accuracy, biasness and predictor selection. The findings also shows that if the predictors are highly or low correlated then its performed more or less same as other methods considering prediction accuracy but again in predictor selection it dominates the other methods. It is catching the attention that in the grouping variables situation proposed method have a commanding position over the other methods. We have also examined its performance in real data application.

The prediction and model complexity have also been demonstrated in *heteroscedastic* regression in

[1] Arthur Bloch

this thesis. We used the idea of unequal inverse variances of error term as weights. We incorporate these weights in penalized regression methods. In the presence of *heteroscedasticity* we have pointed out that the weighted penalized methods significantly reduces both the test errors and model complexity, which was clearly shown in table 1 in chapter 3. The other important finding in chapter 3 is that the proposed method dominates in all simulation examples.

In chapter five we have observed carefully the prediction and regularization performances of five methods in generalized linear models on simulated and real data too. In chapter 5 we display the simulation outcomes in four scenario for logistic regression. The findings shows that the propose and adaptive LASSO both have worse prediction performance. In contrast of this limitation the predictor selection performance is still a little better. In that chapter we have also evaluated the achievement of advertise approaches in Poisson regression. The simulation solutions are outlined in table 4. The prediction performance are almost identical for every method but ridge regression behaves a little better. The proposed approach also does very well the shrinkage and variable selection and there are 4 nonzero variables in the final model. According to oracle results the suggested method perform good. Application to stock market data showed that the proposed technique has minimum classification error and produces most sparse model. The numerical results to micro array data set has also shows that the new technique is a good classification method. To illustrate this we use the leukaemia data consist of 7130 predictors and 72 observations. We have noticed that the proposed classifier gives a 10-fold cross validation error of 1/38 and prediction error of 0/34(which is minimum among all other methods) with 51 predictors(genes) selected. Therefore, we view that the proposed approach is good method(if properly used) for feature selection and prediction in classification problems. In addition, we have calculated the bootstrap standard errors of the simulated mean-squared errors by applying the bootstrap with B=500 re-samplings, to assessing the performance of penalized techniques. The numerical results shows that in most situations the proposed approach has minimum bootstrap standard error.

In this thesis we used 10-fold CV to choose optimum values of the tuning parameters. Everyone model placing and choosing λ were done on training set and prediction or classification error is analyzed on the *holdout* data.

Summing up the above discussion we conclude that the proposed methodology gives good results not only in high-dimensional data but also for $P < n$ situation. The other main conclusion of this thesis is that it more consistent in variable selection than other methods. This implies that it have better *Oracle* properties; and one can say that it performs very good if the accurate model was decided in advance. Finally, our simulation outcomes propound this fresh method into the other L_1 parallel techniques and encourage the implementation in the grouped variables situation.

6.2 Recommendations

We suggest that it will be interesting to further extend this set-up of our technique into threshold models and to see whether one gets improvement in the performance of our regularized method. Second, see that our method has poor performance in logistic regression, it may be further investigated. It is also suggested to investigate the performance of proposed method in high dimensional quantile regression set-up.

Appendix A

Appendix

A.1 Proof of Lemma1

We augment the centered matrix X with P additional rows of, $\sqrt{\lambda}I$ where I is $P \times P$ identity matrix and augment y with zero rows.

The ordinary least squares estimate is given by,

$$\beta^{\tilde{*}}_{OLS} = (y^* - X^*\beta^*)^T(y^* - X^*\beta^*) \tag{A.1}$$

This OLS estimate is based on augmented data set, that is X^* defined as

$$X^* = \begin{bmatrix} X \\ \sqrt{\lambda}I \end{bmatrix}$$

and $y^\star$ is defined as

$$y^\star = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

Hence,

$$\beta^*_{OLS} = (y^*)^T y^* - (y^*)^T X^*\beta^* - (\beta^*)^T(X^*)^T y^* + (\beta^*)^T(X^*)^T X^*\beta^* \tag{A.2}$$

First solving the part,

$$\begin{aligned}(\beta^*)^T(X^*)^T y^* &= \beta^T \begin{bmatrix} X^T & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix} \\ &= \beta^T X^T y \end{aligned} \tag{A.3}$$

Now,

$$\begin{aligned}(y^*)^T X^* \beta^* &= \begin{bmatrix} y^T & 0 \end{bmatrix} \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \beta \\ &= y^T X \beta \end{aligned} \tag{A.4}$$

Now takes the form,

$$\begin{aligned}(X^*\beta^*)^T X^* \beta^* &= \beta^T \begin{bmatrix} X^T & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \\ &= \beta^T X^T X \beta + \lambda I \beta^T \beta \end{aligned} \tag{A.5}$$

Substituting equations(3),(4) and (5) in equation(2), we get,

$$\beta^*_{OLS} = y^T y - y^T X \beta - \beta^T X^T y + \beta^T X^T X \beta + \lambda \beta^T \beta \tag{A.6}$$

We assume orthonormal design matrix, i.e, $X^T X = 1$, then we have

$$\beta^*_{OLS} = y^T y - y^T X \beta - \beta^T X^T y + \beta^T \beta + \lambda \beta^T \beta \tag{A.7}$$

We also know that $(y^T X^T \beta)^T = \beta^T X y$, putting in (7) we get

$$\beta^*_{OLS} = L(\beta) = y^T y - 2 y^T X \beta + \beta^T \beta + \lambda \beta^T \beta \tag{A.8}$$

Now differentiate equation(8), w.r.t β, and equate to 0; we have

$$\frac{\partial L(\beta)}{\partial \beta} = 0 - 2 y^T X + 2\beta + \lambda(2\beta) = 0 \tag{A.9}$$

Or,

$$2\beta - 2 y^T X + 2\lambda\beta = 0 \tag{A.10}$$

Or,

$$2\beta(1+\lambda) = 2y^T X \tag{A.11}$$

Canceling 2 and re-arrange the terms, we get

$$\hat{\beta} = \frac{y^T X}{1+\lambda} \tag{A.12}$$

Clearly, we can see that the resulting estimator is equal to ridge regression, where $y^T X$ is equivalent to OLS estimator based on original data, and λ is threshold parameter of ridge regression.

A.2 Proof of theorem

Given data (y,X), let λ_1 and λ_2 are penalty parameters, and augmented data (y^*, X^*), the proposed method solves a adaptive LASSO type problem(based on augmented data set),

$$\hat{\beta^*} = argmin_{\beta^*}\{|y^* - X^*\beta^*|^2 + \lambda_1 \sum_{j=1}^{p} \hat{w_j}^* |\beta_j^*|\} \tag{A.13}$$

where $\hat{w_j^*} = \frac{1}{|\hat{\beta^*}|}$ are weights. Here $\hat{\beta^*} = \frac{\hat{\beta}^{OLS}}{1+\lambda_2}$ are ridge regression estimators. After simple algebra the above equation become,

$$\hat{\beta^*} = argmin_{\beta}\{|y^* - X^*\beta^*|^2 + \lambda_1(1+\lambda_2) \sum_{j=1}^{p} \hat{w_j} |\beta_j|\} \tag{A.14}$$

Where $\hat{w_j}$ are absolute weights of simple OLS estimator using in adaptive LASSO(H.Zou,2006), λ_2 is tuning parameter of ridge regression.

On augmented data set(y^*, X^*) and original data (y,X) the loss functions are equivalently to each other after some simple matrix algebra. Let $\theta = \lambda_1(1+\lambda_2)$, then last equation(14) become,

$$\hat{\beta^*} = argmin_{\beta}\{|y - X\beta|^2 + \theta \sum_{j=1}^{p} \hat{w_j} |\beta_j|\} \tag{A.15}$$

The last equation is equivalent to adaptive LASSO problem(H.Zou,2006). Conversely of this theorem is also true.

A.3 Proof of Modified LASSO Estimator

Consider the orthogonal design matrix $X_{n\times P}$ and response vector $y_{n\times 1}$. If we centered y_i and standardized X matrix then we get,
$\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} z_i = 0$, and $\sum_{i=1}^{n} z_i^2 = 1$.
The modified adaptive LASSO minimizes the problem,

$$M(\beta, Z, y, \lambda) = \sum_{i=1}^{n}\{y_i - \sum_{j=1}^{P}\beta_j z_{ij}\}^2 + \lambda\sum_{j=1}^{P}\hat{w_j}|\beta_j| \tag{A.16}$$

where $\hat{w_j}$ are suggested weights and z_{ij} are standardized covariates. We can also write the above equation in matrix form.

$$M(.) = (y - Z\beta)^T(y - Z\beta) + \lambda\hat{w_j}|\beta|I_{P\times P} \tag{A.17}$$

were $I_{P\times P}$ is $P \times P$ identity matrix and $|\beta|$ be the diagonal matrix with diagonal entries, for j=1,...,P, is

$$|\beta| = diag(|\beta_j|) = \begin{bmatrix} |\beta_1| & 0 & ... & 0 \\ 0 & |\beta_2| & ... & 0 \\ . & . & ... & . \\ . & . & ... & . \\ 0 & 0 & ... & |\beta_P| \end{bmatrix}$$

We also may write $|\beta|^{-1} = diag(|\beta_j|^{-1})$. Minimizing M, we can get the estimate of β. The equation 2 is form as

$$M(.) = (y^T - \beta^T Z^T)(y - Z\beta) + \lambda\hat{w_j}|\beta|I_{P\times P} \tag{A.18}$$

or,

$$= y^T y - y^T Z\beta - \beta^T Z^T y + \beta^T Z^T Z\beta + \lambda\hat{w_j}|\beta|I_{P\times P} \tag{A.19}$$

or,

$$= y^T y - 2y^T Z\beta + \beta^T Z^T Z\beta + \lambda\hat{w_j}|\beta|I_{P\times P} \tag{A.20}$$

We know that $(y^T Z\beta)^T = \beta^T Z^T y$ Differentiating the last equation w.r.t, β we've,

$$M^{'} = -2y^T Z + 2(Z^T Z)\beta + \lambda \hat{w}_j sign(\beta) \tag{A.21}$$

where,

$$sign(\beta) = \begin{bmatrix} sign(\beta_1) \\ sign(\beta_2) \\ . \\ . \\ . \\ sign(\beta_P) \end{bmatrix}$$

Now assuming the orthonormal columns of Z i.e, $Z^T Z = 1$, equate to 0, and solve for β we get,

$$-2y^T Z + 2\beta + \lambda \hat{w}_j sign(\beta) = 0 \tag{A.22}$$

or,

$$2(-y^T Z + \beta) = -\lambda \hat{w}_j sign(\beta) \tag{A.23}$$

or,

$$\beta^{\hat{ML}} = y^T Z - \frac{\lambda}{2} \hat{w}_j sign(\beta^{\hat{ML}}) \tag{A.24}$$

As we already know that $y^T Z$ is equivalent to OLS estimator in multiple linear regression using matrix algebra. Therefore, the estimator become,

$$\beta^{\hat{ML}} = \hat{\beta}(OLS) - \frac{\lambda}{2} \hat{w}_j sign(\beta^{\hat{ML}}) \tag{A.25}$$

We can observe from last equation that before finding the interested estimators $\beta^{\hat{ML}}$ we need first the signs matrix of that estimators, which is not possible to determine before finding it. For this issue we demonstrate the following Lemma.

According to lemma2, equation 10 is form like,

$$\beta^{\hat{ML}} = \hat{\beta}(OLS) - \frac{\lambda}{2} \hat{w}_j sign(\beta^{\hat{OLS}}) \tag{A.26}$$

Thus, it is the required solution of modified adaptive LASSO mentioned in chapter 2.

A.4 Lemma2

When $y = x - \frac{\alpha}{2} sign(y)$ and $\alpha > 0$, then x and y share the same sign, i.e; $y = x - \frac{\alpha}{2} sign(y) \Rightarrow x = y + \frac{\alpha}{2} sign(y)$. Furthermore, if $y > 0$, it implies that; $y + \frac{\alpha}{2} sign(y) > 0 \Rightarrow x > 0$. Conversely, if y is negative then x is also negative, mathematically, if $y < 0 \Rightarrow x < 0$. Hence, x and y share the same sign.

Chapter 7

Bibliography

[1] Algamal, Z. Y. and Lee, M. H. (2015). Applying penalized binary logistic regression with correlation based elastic net for variables selection. *Journal of Modern Applied Statistical Methods*, 14(1):15.

[2] Amin, M., Song, L., Thorlie, M. A., and Wang, X. (2015). Scad-penalized quantile regression for high-dimensional data analysis and variable selection. *Statistica Neerlandica*, 69(3):212–235.

[3] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

[4] Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132.

[5] Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

[6] Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 311–354.

[7] Daye, Z. J., Chen, J., and Li, H. (2012). High-dimensional heteroscedastic regression with an application to eqtl data analysis. *Biometrics*, 68(1):316–326.

[8] Dong, Y., Song, L., Wang, M., and Xu, Y. (2014). Combined-penalized likelihood estimations with a diverging number of parameters. *Journal of Applied Statistics*, 41(6):1274–1285.

[9] Efron, B. (2000). The bootstrap and modern statistics. *Journal of the American Statistical Association*, 95(452):1293–1296.

[10] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

[11] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

[12] Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

[13] Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). Discussion of boosting papers. In *Ann. Statist.* Citeseer.

[14] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

[15] Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1.

[16] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537.

[17] Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

[18] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

[19] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[20] Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618.

[21] Huber, P. J. (2011). *Robust statistics.* Springer.

[22] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 6. Springer.

[23] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378.

[24] Kwon, S. and Kim, Y. (2012). Large sample properties of the scad-penalized maximum likelihood estimation on high dimensions. *Statistica Sinica*, pages 629–653.

[25] Lee, S., Seo, M. H., and Shin, Y. (2016). The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):193–210.

[26] Lipschutz, S. (1997). *Schaum's outline of theory and problems of beginning linear algebra.* McGraw Hill Professional.

[27] McCullagh, P. and Nelder, J. A. (1989). Generalized linear models, no. 37 in monograph on statistics and applied probability.

[28] MikeWest, C. B., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer using gene expression profiles.

[29] Nevitt, J. and Tam, H. P. (1997). A comparison of robust and nonparametric estimators under the simple linear regression model.

[30] Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34.

[31] Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030.

[32] Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10(6):961–980.

[33] Sun, H. and Wang, S. (2012). Penalized logistic regression for high-dimensional dna methylation data with case-control studies. *Bioinformatics*, 28(10):1368–1375.

[34] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[35] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

[36] Tutz, G. and Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing*, 19(3):239–253.

[37] Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

[38] Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222.

[39] Yang, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

[40] Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703.

[41] Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443.

[42] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

[43] Zou, H. and 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.

[43] [42] [25] [1] [2] [3] [5] [6] [7] [8] [9] [10] [11] [12] [14] [15] [16] [17] [18] [19] [20] [21] [23] [24] [26] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [40] [41] [13] [27] [4] [22] [39]