L_q PENALIZED REGRESSION

by

Yin Xiong

(Under the direction of Cheolwoo Park)

Abstract

For linear regression problems, the Ordinary Least Squares (OLS) model produces unbiased estimates but can have large variances if the design matrix $\boldsymbol{X}$ is close to collinear, and the estimator is not unique if $\boldsymbol{X}$ is less than full rank. To remedy these problems, penalized regression methods such as Ridge, Lasso and Enet have been developed, which have improved OLS in some respects, but failed in others. We study L_q Bridge regressor with $q > 0$ using local quadratic approximation. By letting the q be estimated from the given data, the method extends its practicality. We thoroughly compare all kinds of regression methods under various simulation settings and a real example. Our goal is to assess the performance of L_q regressor and to examine the behavior of tuning parameters. The simulation result shows that L_q regressor is very robust as it performs well in all different cases.

Index words: penalized regression, Ridge, Lasso, Enet, Bridge, L_q penalty

L_q PENALIZED REGRESSION

by

Yin Xiong

M.A., Wuhan University, 1990

M.S., University of Georgia, 2001

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Master of Science

Athens, Georgia

2007

L_q PENALIZED REGRESSION

by

Yin Xiong

Approved:

Major Professor: Cheolwoo Park

Committee: Jeongyoun Ahn
Xiangrong Yin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2007

Acknowledgments

I would like to thank Dr. Mary Meyer, Dr. Xiangrong Yin and Miss Jill Smith who taught me the first statistics courses I've ever taken. Their excellent teaching kindled my interest in statistics and encouraged me to forge ahead toward a new goal of my life.

I would also like to than Dr. Eileen Kraemer and Dr. Maria Hybinette who showed great understanding and support for my taking on a second major while I was doing my doctoral research in Computer Science.

My special thanks go to Dr. Jaxk Reeves who not only taugh me knowledge and techniques in statistics, but also taught me how to write reports and communicate with clients.

It would be impossible to adequately thank my advisor, Dr. Cheolwoo Park, who agreed to be my major professor of master's thesis when he was already very busy in his research and teaching. His profound knowledge in the subject matter and his strong ability to explain complex concepts in an easy-to-understand manner benefitted me enormously. His patient, consistent, and steady guidance is the major factor in the completion of my thesis project.

I would like to thank my advisory committee members, Dr. Jeongyoun Ahn and Dr. Xiangrong Yin, who have provided valuable comments and directions for my research work and in the preparation of this thesis.

Finally, I would like to thank Mr. Youngjoo Yoon who provided the initial R-code for L_q penalized regression methods.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Using standardized variables, consider the linear regression model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \text{for} \quad i = 1, \cdots, n,$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, $\boldsymbol{x}_i^T = (x_{i1}, \cdots, x_{ip})$, and ϵ_i's are independently and identically distributed as normal with mean 0 and variance σ^2.

The vector of coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$ is what a model fitting procedure produces. Typically, the criteria for evaluating the quality of a model include the following:

1. Accuracy of prediction – a model that predicts poorly should not be considered as a good model;

2. Ease of interpretation – a good model should follow the principle of "parsimony" so that the model can better interpret the relationship between the response variable and the predictor variables.

The most common model for linear regression is the ordinary least squares (OLS) model, whose estimates are obtained by minimizing the residual sum of squares (RSS):

$$RSS_{ols} = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2,$$

which yields an estimator

$$\hat{\boldsymbol{\beta}}_{ols} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y},$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ and

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}.$$

The OLS estimator is unbiased, but an unbiased model may still have a large mean-squared-error if it has a large variance. This will be the case if the estimator is highly sensitive to the peculiarities such as noise and the choice of sample points. The OLS estimator is not unique if the design matrix X is less than full rank and the variance is large if X is close to collinear.

The variance can be significantly reduced by deliberately introducing a small amount of bias so that the net effect is a reduction in mean-squared-error. Introducing bias is equivalent to restricting the range of functions for which a model can account.

Penalized regression techniques have been proposed to improve the OLS. Subset selection attempts to improve prediction accuracy by sacrificing a bit of bias in exchange for reducing the variance, thus creating a simpler model for easier interpretation. There are different variants of subset selection. It can be done exhaustively searching all possible subsets. However, even if correct, this approach is not feasible for the large number of predictor variables because the number of subsets grows as 2^p where p is the number of available predictor variables. Subset selection can also be done incrementally, starting with all or none of the available variables, and based on some information criterion remove or add a variable at each step. Subset selection provides interpretable models but can be extremely unstable because it is a discrete process, which means it either drops a variable or retains it. The prediction capability of the model may change significantly if a variable is preserved or discarded.

Continuous version of subset selection is represented by shrinkage methods, also known as regularization. Ridge regression (Hoerl and Kennard, 1970) is an extension of OLS by adding the L_2 penalty term:

$$\boldsymbol{\beta}_{Ridge} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \right\}.$$

Intuitively, the larger the value of λ is, the larger is the shrinkage of the weights. The optimization problem can be equally formalized as

$$\boldsymbol{\beta}_{Ridge} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2, \qquad \text{satisfying} \quad \boldsymbol{\beta}^T \boldsymbol{\beta} \leq s, \text{ for some } s > 0.$$

Ridge regression yields a biased estimator

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\boldsymbol{X}^T + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Ridge regression can be used to solve ill-conditioned linear regression problems. Ill- conditioning means numerical difficulties in performing the matrix inverse necessary to obtain the variance matrix. Since Ridge regression is a continuous process, it is more stable than subset selection. Although Ridge regression shrinks the OLS estimator towards 0 and yields a biased estimator, the variance is smaller than that of OLS estimator. However, Ridge regression does not do variable selection and therefore it does not produce a parsimonious model.

The Lasso (Tibshirani, 1996) is another penalized regression method similar to the Ridge regression but uses the L_1 penalty instead of the L_2 penalty. Lasso estimator is the value that minimizes

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where λ is the penalty parameter. An important feature of Lasso is that it can be used for variable selection because it shrinks some coefficients and sets others to 0, and in this way it tries to retain the good features of both subset selection and Ridge regression. Compared to the classical variable selection methods such as subset selection, the Lasso has two advantages. First, the selection process in Lasso is continuous and hence more stable than the subset selection. Second, the Lasso is computationally feasible for high-dimensional data.

Although the Lasso is quite successful in many situations, it has some limitations of its own. If p, the number of predictor variables, is larger than n, the number of observations, the Lasso selects at most n variables before it saturates because of the nature of the convex

optimization problem. And if the pairwise correlation among a group of variables, then the Lasso tends to select only one variable from the group and does not care which one is selected. To remedy the shortcomings of Ridge and Lasso, Zou and Hastie (2005) proposed a new penalized regression method called elastic net (Enet), which attempt to combine the strength of Ridge and Lasso while being able to deal with group effects and $p > n$ cases.

Frank and Friedman (1993) introduced Bridge regression which minimizes the residual sum of squares subject to a constraint $\sum |\beta_j|^q \leq t$ with $q \geq 0$. It includes Ridge regression with $q = 2$, Lasso regression with $q = 1$, and subset selection with $q = 0$ as special cases. For different values of q, the constrained areas are very different in the parameter spaces. Frank and Friedman did not solve for the estimator of Bridge regression for any given $q > 0$, but they pointed out that it is desirable to optimize the parameter q.

Fu (1998) studied the structure of Bridge estimator and proposed a general approach to solve for the Bridge regression for $q \geq 1$. The shrinkage parameter q and the tuning parameter λ are selected via generalized cross-validation (GCV). Through a simulation study, it is shown that Bridge regression performs well compared to the Lasso and Ridge regression. However, in case of $0 < q < 1$, no computation method is introduced. Knight and Fu (2000) showed some asymptotic properties of the Bridge estimator with $q > 0$. They suggested a computational algorithm for $0 < q < 1$, but no data analysis was done.

We study the Bridge estimator called L_q regression with $q > 0$ using local quadratic approximation. By letting the q be estimated from the given data, the proposed method extends its practicality. When there exist many noise or redundant variables, the Bridge estimator needs variable selection and the q will be estimated to be less than or equal to 1. In other cases, the q will be greater than 1. In a classification setting, Liu *et al.* (2007) used the L_q penalty to support vector machine.

We thoroughly compared OLS, Ridge, Lasso, Enet and L_q regressions under various simulation settings and a real example and found that L_q regression is very robust as it performed well in all these cases.

The thesis is organized as follows: in Chapter 2 we review the literature of penalized regression which includes Lasso, Enet and Bridge regression. In Chapter 3, we elaborate on L_q regression by first giving an brief introduction, then a detailed description of the computation methods. In Chapter 4, we present the simulation results. In Chapter 5, prostate cancer data is analyzed. In Chapter 6 we give summary and suggest future work.

Chapter 2

Literature Review

In this chapter, we introduce the penalized regression methods of Lasso, Enet and Bridge estimator. All the methods are described using the standardized variables. That is,

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \qquad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad \text{for } j = 1, 2, ..., p$$

2.1 Lasso

Lasso (Least absolute shrinkage and selection operator) (Tibshirani, 1996) is a method for estimation in linear models. The Lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant.

Letting $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$, the Lasso estimate $\hat{\boldsymbol{\beta}}$ is defined by

$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\}, \quad \text{subject to } \sum_j |\beta_j| \leq t.$$

The parameter $t > 0$ controls the amount of shrinkage that is applied to the estimates. Let $\hat{\beta}_j^o$ be the full least squares estimates and $t_0 = \sum |\hat{\beta}_j^o|$. Values of $t > 0$ will cause shrinkage of the solutions towards 0, and some coefficients may be exactly equal to 0. For example, if $t = t_0/2$, then the effect will be roughly similar to finding the best subset of size $p/2$.

Lasso is a non-linear and non-differentiable function of the response values even for a fixed value of t. While Ridge regression scales the coefficients by a constant factor, the Lasso translates the coefficients by a constant factor, truncating at 0. Absolute value in Lasso penalty makes the problem of weights estimation non-linear and the penalty tends to drive less important weights to zero faster than the one in Ridge regression.

Consider any model indexed by a vector parameter $\boldsymbol{\beta}$, for which estimation is carried out by maximization of a function $l(\boldsymbol{\beta})$. To apply the Lasso, we maximize $l(\boldsymbol{\beta})$ under the constraint $\Sigma|\beta_j| \leq t$. It might be possible to carry out this maximization by a general (non-quadratic) programming procedure, but here Tibshirani considered models for which a quadratic approximation to $l(\boldsymbol{\beta})$ leads to an iteratively reweighted least squares (IRLS) procedure for computation of $\boldsymbol{\beta}$. Using this approach, the constrained problem is solved by iterative application of the Lasso algorithm, within a IRLS loop.

Tibshirani (1996) provided three methods for the estimation of the Lasso parameter t: cross-validation, generalized cross-validation and an analytical unbiased estimate of risk.

The Lasso can be applied to many other models, for example, the proportional hazards model. The Lasso can also be applied to generalized regression models (Klinger, 2001).

2.2 Elastic Net

Elastic Net (Zou and Hastie, 2005) is a regularization and variable selection method. Elastic Net attemps to remedy the shortcomings of Lasso, namely the poor performance in $p > n$ cases and when pairwise correlations exist among a group of variables. These two situations can arise in gene selection problems. For example, a typical microarray data set has many thousands of predictors (genes) but often fewer than several hundred of samples. For those genes sharing the same biological "pathway", the correlations between them can be high (Segal and Conklin, 2003). Apparently, Lasso is not a proper method to deal with such cases and Elastic Net targeted to find a new method that retain the advantages of Lasso and at the same time remedies the disadvantages of it.

The naive elastic net criterion for any fixed non-negative λ_1 and λ_2 is defined as

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|^2 + \lambda_2|\boldsymbol{\beta}|^2 + \lambda_1|\boldsymbol{\beta}|_1,$$

where

$$|\boldsymbol{\beta}|^2 = \sum_{j=1}^{p} \beta_j^2,$$

$$|\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j|.$$

The naive elastic net estimator $\hat{\boldsymbol{\beta}}$ is the minimizer of $L(\lambda_1, \lambda_2, \boldsymbol{\beta})$.

This procedure can be viewed as a penalized least squares method. Let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, then solving $\hat{\boldsymbol{\beta}}$ is equivalent to the optimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|^2, \quad \text{suject to} \quad (1-\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2 \leq t \quad \text{for some} \quad t.$$

The function $(1-\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2$ is called the elastic net penalty, which is a convex combination of the Lasso and Ridge penalty. When $\alpha = 1$, the naive elastic net becomes simple Ridge regression. For all $\alpha \in [0, 1)$, the elastic net penalty function is singular (without first derivative) at 0 and is strictly convex for all $\alpha > 0$, thus having the characteristics of both Lasso and Ridge regression.

The naive elastic net estimator is a two-stage procedure: for each fixed λ_2, the Ridge regression coefficients are found first, and then the Lasso-type shrinkage is done along the Lasso coefficient paths.

To find the elastic net estimator, define an artificial data set $(\boldsymbol{y}^*, \boldsymbol{X}^*)$ by

$$\boldsymbol{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2}(\boldsymbol{X}, \sqrt{\lambda}_2\boldsymbol{I})^T, \quad \boldsymbol{y}^*_{(n+p)} = (\boldsymbol{y}, \boldsymbol{0})^T$$

Let $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$ and $\boldsymbol{\beta}^* = \sqrt{(1+\lambda_2)}\boldsymbol{\beta}$, then the naive elastic net criterion can be written as

$$L(\gamma, \boldsymbol{\beta}) = L(\gamma, \boldsymbol{\beta}^*) = |\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*|^2 + \gamma|\boldsymbol{\beta}^*|_1.$$

Let

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} L(\gamma, \boldsymbol{\beta}^*)$$

Then

$$\hat{\boldsymbol{\beta}}_{Enet} = \frac{1}{\sqrt{(1+\lambda_2)}}\hat{\boldsymbol{\beta}}^*.$$

2.3 Bridge Regression

Bridge regression minimizes $RSS = \sum(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$ subject to a constraint $\sum |\beta_j|^q \leq t$ with $q > 0$. The Bridge estimator can be obtained equivalently by

$$\hat{\boldsymbol{\beta}}_{Bridge} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}.$$

Figure 2.1: Contours of constant value of $\sum_j |\beta_j|^q$ for given values of q

Figure 2.1 depicts the contours of constant value of $\sum_j |\beta_j|^q$ for given values of q in two dimensions. Subset selection corresponds to $q = 0$. Lasso regression corresponds to $q = 1$ which has the advantage of being closer to subset selection than is Ridge regression ($q = 2$) and is also the smallest value of q giving a convex region.

Fu (1998) studied the structure of the Bridge estimators and developed a general approach to solve Bridge regression for $q \geq 1$.

Let $G(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}, \lambda, q) = RSS + \lambda \sum |\beta_j|^q$. G is convex in $\boldsymbol{\beta}$, and $G \to +\infty$ as $||\boldsymbol{\beta}|| \to +\infty$. Thus function G can be minimized. There exists a $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} G(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}, \lambda, q)$. Take partial derivative of G with respect to β_j at $\beta_j \neq 0, j = 1, \cdots, p$. Denote $S_j(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{\partial RSS}{\partial \beta_j}$ and $d(\beta_j, \lambda, q) = \lambda q |\beta_j|^{q-1}$ sign (β_j). Setting $\frac{\partial G}{\partial \beta_j} = 0$ leads to

$$\begin{cases} S_1(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}) + d(\beta_1, \lambda, q) &= 0 \\ \quad \vdots & \\ S_p(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}) + d(\beta_p, \lambda, q) &= 0. \end{cases} \qquad (P1)$$

Denote $\boldsymbol{\beta}$ by $(\beta_j, \boldsymbol{\beta}^{-j})^T$, where $\boldsymbol{\beta}^{-j}$ is a $(p-1)$ vector consisting of the β_i's other than β_j. The jth equation of (P1) is:

$$S_j(\beta_j, \boldsymbol{\beta}^{-j}, \boldsymbol{X}, \boldsymbol{y}) = -d(\beta_j, \lambda, q). \tag{2.1}$$

The left hand side function of (2.1), LHS $= 2\boldsymbol{x}_j^T\boldsymbol{x}_j\beta_j + \sum_{i\neq j} 2\boldsymbol{x}_j^T\boldsymbol{x}_j\beta_j - 2\boldsymbol{x}_j^T\boldsymbol{y}$, is, for fixed $\boldsymbol{\beta}^{-j}$, a linear function of β_j with positive slope $2\boldsymbol{x}_j^T\boldsymbol{x}_j$. The right hand side function of (2.1), RHS $= -\lambda q|\beta_j|^{q-1}$, is nonlinear in β_j. RHS is of different shape for different value of q. It is continuous, differentiable, and monotonically decreasing for $q > 1$ except nondifferentiable at $\beta_j = 0$ for $1 < q < 2$, a heavy-side function with a jump of height 2λ at $\beta_j = 0$ for $q = 1$. Therefore, equation (2.1) has a unique solution for $q > 1$, a unique solution or no solution for $q = 1$. Here we assume $q > 1$.

To compute the Bridge estimator for $q > 1$, the modified Newton-Raphson method was proposed. Here is the Modified Newton-Raphson (M-N-R) Algorithm for the Bridge regression with $q > 1$:

1. Start with an initial value $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_{ols} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$.

2. At step m, for each $j = 1, ..., p$, let $S_0 = S_j(0, \hat{\boldsymbol{\beta}}_{-j}, \boldsymbol{X}, \boldsymbol{y})$. Set $\hat{\boldsymbol{\beta}} = 0$ if $S_0 = 0$. Otherwise, if $q \geq 2$, apply the Newton-Raphson method to solve for the unique solution $\hat{\beta}_j$ of equation (2.1); if $q < 2$, modify function $-d$ by changing one part to its tangent line at some point between the solution and the origin. Then apply the Newton-Raphson method to equation (2.1) with the modified function $-d$ to solve for the unique solution $\hat{\beta}_j$. Form a new estimator $\hat{\boldsymbol{\beta}}_m = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$ after updating all $\hat{\beta}_j$.

3. Repeat 2 until $\hat{\boldsymbol{\beta}}_m$ converges.

Bridge regression is a special family of penalized regressions with two very important members – Ridge regression and Lasso regression. It plays an important role in solving collinearity problems. It yields a small variance of the estimator and achieves good estimation and prediction by shrinking the estimator towards 0.

Chapter 3

L_q Regression

3.1 Introduction

To achieve the purpose of variable selection, certain conditions can be set on the penalty functions. Antoniadis and Fan (2001) and Fan and Li (2001) argued that a good penalty function should result in an estimator with the following three properties:

1. unbiasedness for a large true coefficient to avoid excessive estimation bias;

2. sparsity (estimating a small coefficient as zero) to reduce model complexity; and

3. continuity to avoid unnecessary variation in model prediction.

Usually, however, the penalty functions such as L_1 and L_2 do not simultaneously satisfy these three mathematical conditions, which require that $p_\lambda(\cdot)$ should be a concave function over $(0, 1)$ and that $p'_\lambda(0+) > 0$. The latter condition is referred to as singularity at the origin. Fan and Li (2001) suggested the SCAD (Smoothly Clipped Absolute Deviation) which satisfies all the three conditions. Also, they suggested that L_q penalty with $q < 1$ might satisfy them.

In practice, a learning procedure of L_q penalty with a fixed q has its advantages over others only under certain situations, because different types of penalties may suit best for different data structures. Since the best choice of q varies from problem to problem, we propose to treat q as a tuning parameter and select it adaptively.

The simple and special structure of the Bridge estimator for $q > 1$ makes the computation very simple. The Bridge operator is nonlinear for $q \neq 2$. The nonlinearity of the Bridge

operator makes it perform very differently from the Ridge operator ($q = 2$) or the OLS operator ($\lambda = 0$). It is not a surprise that the Bridge model does not always perform the best in the estimation and prediction compared to the other shrinkage models. Therefore, new optimization techniques are desirable, especially for nonlinear operators.

In this thesis, we study an algorithm for the Bridge estimator with $q > 0$ using local quadratic approximation. By letting the q be estimated from the given data, the method extends its practicality. When there exist many noisy or redundant variables, the Bridge estimator needs variable selection and the q will be estimated to be less than or equal to 1. In other cases, the q will be greater than 1.

3.2 Computation

If $0 < q < 1$, the minimization problem is not convex and it needs to be solved differently. Fan and Li (2001) proposed an algorithm for minimizing penalized general loss via local quadratic approximations. Denote a loss function by $l(\boldsymbol{\beta})$. Then the penalized general loss can be written in a unified form as

$$l(\boldsymbol{\beta}) + n\sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{3.1}$$

Under some mild conditions, the penalty term can be locally approximated at $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p})^T$ by a quadratic function:

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{0j}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_{0j}|)}{|\beta_{0j}|}(\beta_j^2 - \beta_{0j}^2).$$

In the L_q case, $p_\lambda(|\beta_j|) = \lambda|\beta_j|^q$ and $p'_\lambda(|\beta_j|) = \lambda q|\beta_j|^{q-1}$.

Therefore the minimization problem of the unified form can be reduced to a quadratic minimization problem and the Newton-Raphson algorithm can be used. The solution for the penalized least squares problem can be found by iteratively computing the Ridge regression (Fan and Li, 2001). Minimizing (3.1) is equivalent to the quadratic approximation. Figure 3.1 summarizes the proposed unified algorithm.

$\boldsymbol{y}$ is an $n \times 1$ vector and $\boldsymbol{X}$ is an $n \times p$ matrix.
For given q, λ
(a) Initialize $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p})^T$.
(b) Until ($\hat{\boldsymbol{\beta}}_j$ is converged)

$$\hat{\boldsymbol{\beta}}_j = \left\{ \boldsymbol{X}^T \boldsymbol{X} + n\boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}}_{j-1}) \right\}^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

where $\boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}}_{j-1}) = \text{diag}(\lambda q |\hat{\beta}_{j-1,1}|^{q-2}, \cdots, \lambda q |\hat{\beta}_{j-1,p}|^{q-2})$

Figure 3.1: Computation method of the Bridge estimator

For L_q regression, there are two tuning parameters λ and q. The parameter λ controls the tradeoff between minimizing the loss and the penalty. The other tuning parameter q determines the penalty function. The proper choice of q is important and depends on the nature of data. If there are many noise input variables, the L_q penalty with $q \leq 1$ is desired since it automatically selects important variables. On the other hand, if all the covariates are important, it may be more preferable to use $q > 1$ to avoid unnecessary variable deletion. Therefore, q should be chosen adaptively by data.

L_q finds the optimal conbination of λ and q by grid-search which is a brute-force search method or exhaustive search method, thus the L_q algorithm searches the whole parameter space which is a $l \times m$ grid where l is the number of the tuning parameter q and m is the number of the parameter λ. The optimal solution is the combination of q and λ that minimizes the penalty function. The advantage of grid search is that it seaches the whole parameter space and is guaranteed to find the optimal solution if such a solution exists. But when the search space is very large, the computation could be so time-consuming that it loses practical meaning.

We introduce two more algorithms for solving (3.1), namely, L_q2 and L_qFu. L_q2 is a variation of L_q and makes use of Ridge regression in its approximation process.

L_q2 begins with the objective function of L_q

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^{p}\frac{\lambda q}{2}|\beta_{0j}|^{q-2}\beta_j^2.$$

Given q, let $\gamma_j = \frac{q}{2}|\beta_{0j}|^{q-2}$, and

$$x_{ij}^* = x_{ij}/\sqrt{\gamma_j}, \quad \boldsymbol{\beta}_j^* = \boldsymbol{\beta}_j\sqrt{\gamma_j}.$$

The loss function can be written as

$$\sum(y_i - \Sigma\beta_j x_{ij})^2 = \sum(y_i - \Sigma\beta_j\sqrt{\gamma_j}\cdot x_{ij}/\sqrt{\gamma_j})^2 = \sum(y_i - \Sigma\beta_j^* x_{ij}^*)^2,$$

and

$$n\lambda\sum\frac{q}{2}|\beta_{0j}|^{q-2}\beta_j^2 = n\lambda\Sigma\gamma_j\beta_j^2 = n\lambda\Sigma\gamma_j\beta_j^{*2}/\gamma_j = n\lambda\Sigma\beta_j^{*2}.$$

The above transformatin gives us the one-tuning-parameter version of the loss function:

$$\sum(y_i - \boldsymbol{x}_i^{*T}\boldsymbol{\beta}^*) + n\lambda\boldsymbol{\Sigma}\beta_j^{*2}. \tag{3.2}$$

Given q, we find $(\boldsymbol{\beta}^*, \lambda)$ which minimizes (3.2) using Ridge regression. After trying all the q's, we find the optimal solution $(\boldsymbol{\beta}^*, \lambda, q)$ and the original $\hat{\boldsymbol{\beta}}$ can be obtained by the reverse transformation.

L_qFu is a combination of L_q local quadratic approximation version and the Modified Newton-Raphson (M-N-R) algorithm proposed by Fu (1998). For cases where $q > 1$, the Modified Newton-Raphson algorithm in Section 2.3 is used. For cases when $q \leq 1$, local quadratic approximation described above is used. The reason we use this combination is that when $q > 1$, the minimization problem becomes convex and it does not need the quadratic approximation.

Chapter 4

Simulation

In order to find out how good our L_q estimators are in comparison to other penalized regression techniques, we performed a simulation study using Ordinary Lease Square regression (OLS), Ridge regression, Lasso, Elastic net (Enet), L_q, $L_q 2$ and $L_q Fu$.

4.1 Simulation Settings

The simulation has four settings. The first three were used in the original Lasso paper (Tibshirani, 1996). Their major concern is to compare the prediction performance of Lasso and Ridge regression systematically. The fourth setting was originally presented in the Enet paper (Zou and Hastie, 2005). Its major concern is to create a grouped variable situation and compare the performance between Lasso and Enet.

For each setting, our simulated data consist of three data sets: a training set, an independent tuning set and an independent test set. The training data were used to fit the model, and the tuning data were used to select the tuning parameters. The specific settings are as follows:

1. Setting 1: 50 simulated data sets, each consisting of 20/20/200 (training/tuning/test) observations and 8 predictors with $\beta = (3, 1.5, 0, 0.2, 0, 0, 0)$ and $\sigma = 3$. The pairwise correlation between x_i and x_j was set to be $corr(i, j) = 0.5^{|i-j|}$.

2. Setting 2: the same as setting 1, except that $\beta_j = 0.85$ for all $j's$.

3. Setting 3: 50 simulated data sets, each consisting of 100/100/400 observations and 40 predictors. The predictors were set to be:

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

and $\sigma = 15$; $corr(i, j) = 0.5$ for all i and j.

4. Setting 4: 50 simulated data sets, each consisting of 50/50/400 observations and 40 preditors. The predictors were set to be:

$$\beta = (\underbrace{3, \cdots, 3}_{15}, \underbrace{0, \cdots, 0}_{25})$$

and $\sigma = 15$. The predictors X were generated as follows:

$$\begin{aligned} x_i &= Z_1 + \epsilon_i^x, \quad Z_1 \sim N(0,1), \quad i = 1, \cdots, 5, \\ x_i &= Z_2 + \epsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 6, \cdots, 10, \\ x_i &= Z_3 + \epsilon_i^x, \quad Z_3 \sim N(0,1), \quad i = 11, \cdots, 15, \\ x_i &\sim N(0,1), iid, i = 16, \cdots, 40. \end{aligned}$$

where ϵ_i^x are errors in relation to the x and they are independent, identically distributed $N(0, 0.01), i = 1, \cdots, 15$. In this model, we have three equally important groups, and within each group there are five members. There are also 25 pure noise features. An ideal method would select only the 15 true features and set the coefficients of the 25 noise features to 0.

We used both MSE and PSE to measure the simulation results. For simulations where variable selection is the main issue, we also take the number of excluded β's into consideration. Suppose that

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $E(\epsilon)$=0 and var(ϵ)=σ^2. The mean-squared error of an estimate $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is defined by

Table 4.1: Estimated Model Errors

Sim	MSE	OLS	Ridge	Lasso	Enet	L_q	L_q2	L_qFu
1	Mean	5.85	3.60	3.09	3.06	3.23	3.25	3.17
	s.e	0.47	0.34	0.31	0.31	0.33	0.34	0.30
	Median	5.03	2.78	2.68	2.51	2.76	2.76	2.78
2	Mean	7.28	2.38	3.96	3.15	2.57	2.59	2.81
	s.e	0.79	0.25	0.31	0.22	0.31	0.31	0.32
	Median	14.48	2.03	3.25	2.78	1.85	1.93	2.08
3	Mean	148.01	24.26	47.68	34.7	26.14	26.14	27.00
	s.e	6.75	1.03	1.66	1.05	1.57	1.57	1.59
	Median	143.31	22.71	47.66	34.96	24.06	24.06	24.06
4	Mean	1060.48	73.34	86.81	29.63	34.15	32.94	32.10
	s.e	93.10	4.38	7.67	3.89	4.58	4.48	4.66
	Median	892.48	67.41	71.78	20.47	23.11	20.89	18.52

$$MSE = E(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})^2.$$

A similar measure is the prediction error of $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ given by

$$PSE = E(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2 = MSE + \sigma^2.$$

For all the four simulation settings, we set the same search space for the tuning parameters λ's and q's. The choices of λ are: $\lambda_k = 2^{k-6}$, $k = 1, \cdots, 20$. The possible choices of q are: $(0.1, 0.4, 0.7, 1, 1.3, 1.7, 2, 2.5, 3)$.

4.2 Simulation Results

Both Table 4.1 and Table 4.2 show the result of our simulation. The L_q method is the second best in most cases and this shows that it is very robust in various circumstances. It may not

Table 4.2: Estimated Prediction Error

Sim	PSE	OLS	Ridge	Lasso	Enet	L_q	L_q2	L_qFu
1	Mean	14.79	12.59	12.15	11.93	12.25	12.21	12.37
	s.e	0.47	0.38	0.41	0.36	0.41	0.40	0.43
	Median	13.85	11.95	11.45	11.48	11.46	11.45	11.66
2	Mean	16.40	11.54	12.83	12.10	11.71	11.66	11.88
	s.e	0.93	0.31	0.31	0.31	0.33	0.34	0.35
	Median	14.48	11.24	12.66	11.83	11.22	11.27	11.32
3	Mean	372.96	250.96	273.99	260.90	252.77	252.77	254.25
	s.e	6.90	2.57	2.74	2.75	2.87	2.87	2.94
	Median	366.00	249.48	270.93	258.87	250.66	250.66	253.72
4	Mean	1285.75	298.56	310.20	253.83	258.91	260.78	256.85
	s.e	93.89	5.22	8.02	3.71	4.73	5.16	4.59
	Median	1182.60	294.12	302.46	251.10	253.23	254.38	254.26

be the best performer for a single simulation setting, but it outperformed the others in one way or another. It is always close to the best no matter what particular structures the data sets possess.

The first simulation setting has 8 predictor variables and the true values of 5 out of 8 are 0. In this case, Lasso is supposed to have the best performance because Lasso does variable selection. From the simulation result we can see that Enet performed the best, Lasso came the second, and L_q is also very close to the performance of Lasso. The mean values of the optimal q's are very close to 1, with 1.05, 1.05 and 0.98 for L_q, L_q2 and L_qFu respectively.

The second simulation setting also has 8 predictor variables, but this time the true values of all the 8 β's are 0.85, which means no variable selection is intended. Since there exist correlation structures, Ridge regresion ($q = 2$) is expected to perform the best. It did perform the best, with L_q closely followed. The mean MSE for Ridge and the L_qs are 2.38,

2.57, 2.59 and 2.81 respectively while the mean MSE for Lasso and Enet are both above 3. The mean values for the optimal q's selected by L_q, L_q2 and L_qFu are 2.30, 2.33 and 1.72 respectively, close to Ridge regression ($q = 2$).

The third simulation setting has 40 predictor variables and there is high collinearity among these variables. Also there are 20 zeros out of the 40 coefficients. Thus the algorithm should handle both sparsity and collinearity. Ridge regression performs best in this case while both Lasso and Enet failed to handle collinearity in a satisfactory way. It is interesting to see that Enet performed poorly because it is a combination of Lasso and Ridge in some sense. L_q is very close to Ridge in this case. The mean MSE for Ridge and the L_qs are all within the range of 20's while the mean MSE for Lasso and Enet are in the 40's and 30's respectively. The mean value of the optimal q's selected by L_q, L_q2 and L_qFu are 2.38, 2.38 and 1.85 respectively, very close to Ridge regression ($q = 2$).

The fourth simulation setting is specifically suitable for Enet because it deals with group effect and Enet was designed with such effect in mind. Also, 15 out of 40 coefficients are zeros. As expected, Enet outperforms all the others, but L_q follows closely. For this simulation setting, the mean values of the optimal q's selected by L_q, L_q2 and L_qFu are 0.53, 0.54 and 0.40 respectively.

The simulation results suggest that when the structures of the data set is not known beforehand, L_q is the best choice for penalized regression. Since it can adapt to the data and handle most of the cases in a satisfactory manner.

Figures 4.1, 4.2, 4.3, and 4.4 plot the optimal q's against the optimal λ's for the 50 runs for Simulations 1, 2, 3, 4 respectively. The dots are usually less than 50 because some of them are overlapped with each other. The figures visually tell us the relationship between λ's and q's. For all the 4 simulations, most of the optimal combination of λ and q lie in either the left or the lower-left corner. The combination of large λ and large q never happened.

Table 4.3 shows the mean, median and standard deviation of the optimal q's selected by the three L_q methods for the four simulation settings. Among the 8 predictor variables for

Table 4.3: Optimal q

Sim	q	L_q	L_q2	L_qFu
1	Mean	1.05	1.05	0.98
	s.e	1.01	1.01	0.84
	Median	0.85	0.07	0.85
2	Mean	2.30	2.33	1.72
	s.e	0.94	0.91	0.56
	Median	3.00	3.00	2.00
3	Mean	2.381	2.38	1.85
	s.e	0.72	0.72	0.37
	Median	2.50	2.50	2.00
4	Mean	0.53	0.54	0.398
	s.e	0.80	0.80	0.45
	Median	0.40	0.10	0.10

the Simulation 1, 5 of them are 0's, thus we expect the optimal q to be less than or equal to 1, which does the variable selection. As the table shows, all three L_q methods produced the mean q value around 1. Since the number of zeros is not large, the estimated q was not much less than 1. For the Simulation 2, the true values of the 8 β's are all 0.85, so variable selection is not expected here and we see that the optimal q's are mostly around 2. The 40 predictor variables for the Simulation 3 are highly correlated and 20 of them are 0's. Interestingly, most of the q value are very close to 2, which means that L_q chose to take care of collinearity rather than variable selection. As for the Simulation 4, there exist group effects and sparsity (25 of 40 coefficients are zeros) and q tends to be less than 1. In this case, L_q takes care of variable selection.

Table 4.4 contains the mean, median and standard deviation of the optimal λ's selected by Ridge, Lasso, Enet and L_q for the four simulation settings we described in previous section.

Table 4.4: Optimal λ

Sim	λ	Ridge	Lasso	Enet(λ1)	Enet(λ2)	L_q	L_q2	L_qFu
1	Mean	9.23	0.66	329.86	0.68	1.67	3.81	2.37
	s.e	13.88	0.21	2316.74	0.25	2.91	7.07	4.02
	Median	4 .00	0.65	0.25	0.70	0.50	1.00	1.00
2	Mean	13.11	0.66	1645.49	0.69	0.59	1.03	0.90
	s.e	9.74	0.23	4962.76	0.25	1.22	1.54	1.23
	Median	8.00	0.65	1.00	0.70	0.50	1.00	0.50
3	Mean	128.0	0.43	5.00	0.38	0.65	1.31	1.41
	s.e	77.58	0.07	6.70	0.20	0.93	1.66	0.92
	Median	128.0	0.40	2.00	0.40	0.50	0.50	1.00
4	Mean	0.23	0.66	329.86	0.46	1.67	3.81	2.38
	s.e	13.88	0.21	2316.74	0.10	2.91	7.08	4.02
	Median	4.00	0.65	0.25	0.50	0.50	1.00	1.00

Table 4.5: Variable selection: number of excluded β's (out of 50)

Order	True value of β	Lasso	Enet	L_q	L_q2	L_qFu
1	3	0	0	0	0	0
2	1.5	0	3	6	2	0
3	0	0	15	26	12	0
4	0	1	13	28	14	1
5	2	0	3	15	2	0
6	0	0	20	32	15	0
7	0	1	25	38	23	1
8	0	2	26	39	24	2
Right Exclusion		4(1.6%)	99(39.6%)	163(65.2%)	88(35.2%)	4(1.6%)
Wrong Exclusion		0(0%)	6(4%)	21(14%)	4(2.7%)	0(0%)

Table 4.6: A comparison of running times between L_q, L_q2 and L_qFu

method	Simulation 1	Simulation 2	Simulation 3	Simulation 4
L_q	3:45	4:01	19:21	9:44
L_q2	5:14	5:50	27:45	25:06
L_qFu	7:10	6:31	1:06:51	50:58

Table 4.5 contains the number of β's that were set to 0 in Simulation 1 by Lasso, Enet, L_q, L_q2 and L_qFu. For the 8 β's, only β_1, β_2 and β_5 are non-zero. All the rest are zero's, so variable selection is expected here. The result shows that Lasso and L_qFu failed to do variable selection as both of them only correctly set 1.6% of the β's to zero. L_q set more β's to zero (163 correctly and 21 incorrectly) than any other methods.

We do not report the number of β's set to 0 for Simulations 2, 3, and 4 since they are either not intended for variable selection or combined with other issues that need to be taken care of. For example, Simulation 3 not only has half of its true β's equal to zero, but also involves high collinearity among the predictor variables. Simulation 4 exhibits group effects at the same time when a large part of the β's are zero's. When collinearity or group effect is mixed with variable selection, none of the methods dealt with both successfully.

As for the three L_q methods, though their performance is similar to each other, their running time differs very much. Table 4.6 contains the running times of L_q, L_q2 and L_qFu for Simulations 1, 2, 3 and 4. These are average running times over three runs for each on a personal computer. It turned out that L_q runs the fastest while L_qFu runs the slowest, with L_q2 in the middle.

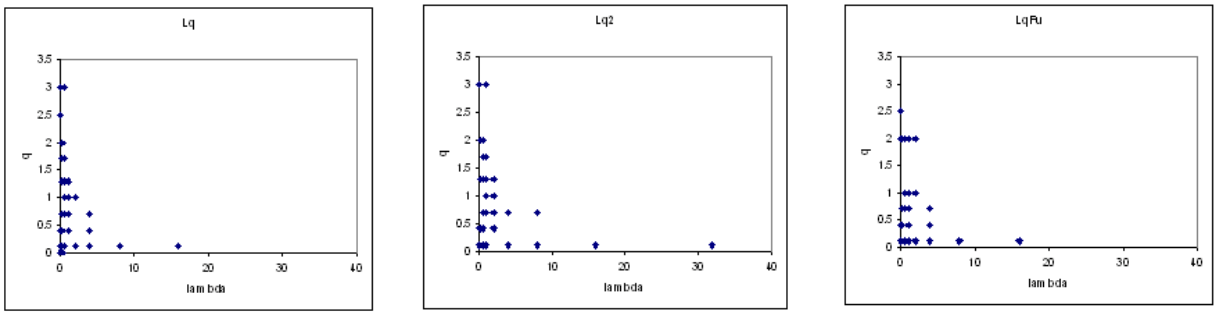

Figure 4.1: Distributions of λ and q for 50 runs of simulation 1

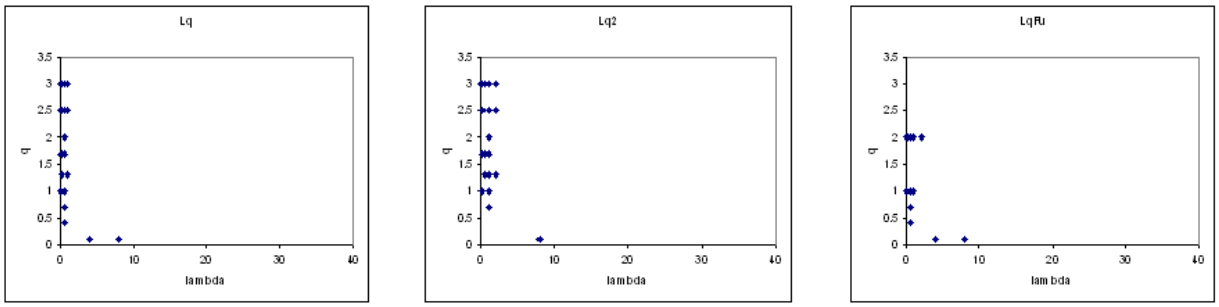

Figure 4.2: Distributions of λ and q for 50 runs of simulation 2

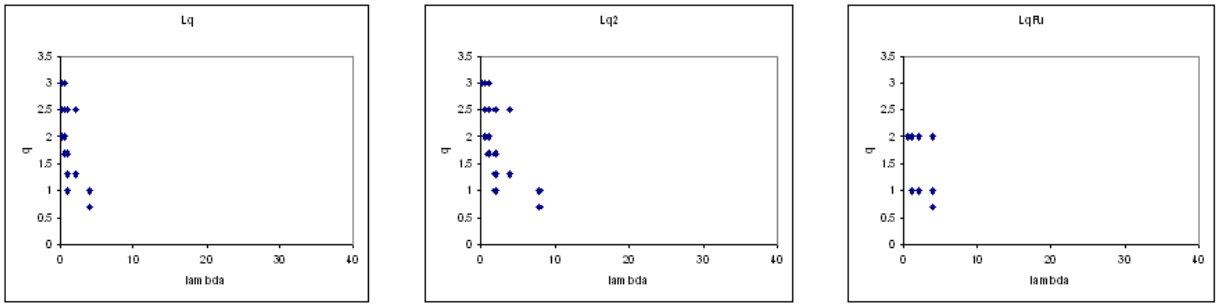

Figure 4.3: Distributions of λ and q for 50 runs of simulation 3

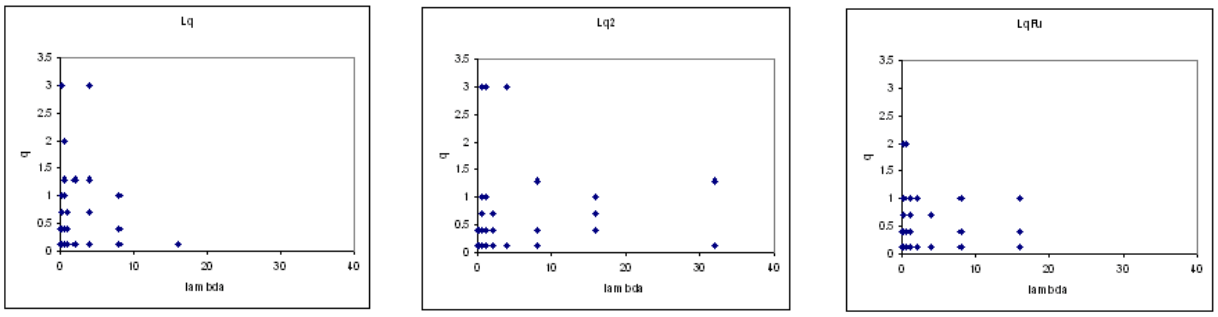

Figure 4.4: Distributions of λ and q for 50 runs of simulation 4

Chapter 5

Prostate Cancer Data Analysis

The prostate cancer data come from a study by Stamey et al. (1989) that examined the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The study had a total of 97 observations of male patients aged from 41 to 79 years. The response variable is *lpsa* - the logarithm of prostate-specific antigen. The covariates are as follows:

1. lcavol - log (cancer volume)
2. lweight - log (prostate weight)
3. age
4. lbph - log (benign prostatic hyperplasia amount)
5. svi - seminal vesicle invasion
6. lcp - log (capsular pEnetration)
7. gleason - Gleason score
8. pgg45 - percentage Gleason scores 4 or 5

As Table 5.1 shows, certain correlation is present between the covariates. The pairwise coefficient is 0.752 between gleason and pgg45, 0.673 between svi and lcp, and 0.675 between lcavol and lcp, and so on. The condition number is 16.9, which indicates a slight collinearity in the covariates.

Table 5.1: Correlatioin Matrix of the Covariates of the Prostate Cancer Data

lcavol	1.000	0.194	0.225	0.027	0.539	0.674	0.432	0.434
lweight	0.194	1.000	0.308	0.435	0.109	0.100	-0.001	0.051
age	0.225	0.308	1.000	0.350	0.118	0.128	0.269	0.276
lbph	0.027	0.435	0.350	1.000	-0.086	-0.007	0.078	0.078
svi	0.539	0.109	0.118	-0.086	1.000	0.673	0.320	0.458
lcp	0.675	0.100	0.128	-0.007	0.673	1.000	0.515	0.632
gleason	0.432	-0.001	0.269	0.078	0.320	0.515	1.000	0.752
pgg45	0.434	0.051	0.276	0.078	0.458	0.632	0.752	1.000

Table 5.2: Prostate cancer data: comparing different methods

Method	Parameter(s)	MSE	Variables selected
OLS		0.5863	All
Ridge	$\lambda = 2$	0.5709	All
Lasso	$\lambda = 0.9$	0.5863	All
Enet	$\lambda_1 = 0.03$, $\lambda_2 = 1$	0.5725	7
L_q	$\lambda = 0.03$, $q = 2$	0.5606	7
$L_q 2$	$\lambda = 0.03$, $q = 0.1$	0.5694	7
$L_q Fu$	$\lambda = 0.03$, $q = 0.5$	0.5572	7

OLS, Ridge regression, Lasso, Enet, L_q, $L_q 2$ and $L_q Fu$ were all applied to these data. The 97 observations was divided into two parts: a training set with 67 observations and a test set with 30 observations. Model fitting and tuning parameter selection by 10-fold CV were carried out on the training data. We then compared the performance of these methods by computing their prediction mean-square error on the test data.

Table 5.2 shows that while all the methods performs similarly, the L_q methods have the lowest PSEs. The $L_q Fu$ has the lowest PSE, 0.5572. Enet, L_q, $L_q 2$, and $L_q Fu$ identified

gleason as an insignificant variable and excluded it from the predictor variables. The q values were estimated as 2, 0.1, and 0.5 respectively for L_q, $L_q 2$ and $L_q Fu$.

L_q methods selected variables and the q's were estimated less than 1 except L_q. For some reason, L_q did not estimate q less than 1. Further investigation is needed to explain this phenomenon.

Chapter 6

Summary

In this thesis we studied L_q Bridge method for shrinkage and selection for regression and generalized regression problems. We have presented some evidence that suggests that the L_q estimator with $q > 0$ using local quadratic approximation is a worthy competitor to Ridge, Lasso and Enet regression methods. We examined the relative merits of the methods in four different scenarios:

1. variable selection – Enet and Lasso do the best here, L_q closely follows, Ridge regression doesn't do very well;

2. correlation beween variables – Ridge does the best, L_q closely follows, Lasso does quite poorly, Enet is somewhere midway between L_q and Lasso;

3. high collineairty and variable selection – Ridge does the best, L_q closely follows, Enet comes the third while Lasso does quite poorly;

4. group effect and variable selection – Enet does the best as designed, L_q closely follows, Ridge and Lasso do poorly with their mean MSE almost having doubled that for L_q.

L_q penalized regression is a generalization of Ridge regression and subset selection, through the addition of a penalty of the form $\lambda \sum_j |\beta_j|^q \leq t$. The Lasso corresponds to $q = 1$ and Ridge corresponds to $q = 2$. The simulation result shows that L_q regression with the joint estimation of the λ and q is a very robust penalized regression method. It can adapt to different features of the data and perform well in various situations. When the features of the data is not known beforehand, it is safe to use L_q regression because it guarantees

to produce at least the second best result, very close to the regression method designed specifically for a certain situation.

The encouraging results reported in this thesis suggest that L_q penalized regression might prove to be useful in a wide variety of statistical estimation problems. Further study is needed to investigate these possibilities. In L_q regression, we found optimal λ and q by grid-search. Since this is time-consuming, it is desirable to develop a faster algorithm to find the optimal solution.

In simulation setting 3, we could observe that when there exist both high correlation between variables and noise variables needed to be removed, L_q tends to take care of collinearity since q was estimated around 2. This observation needs further investigation by controlling the degree of correlatioin and the number of noise variables.

Bibliography

[1] Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussion). *Journal of American Statistical Association*, **96**, 939-967.

[2] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

[3] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, **35**, 109-148.

[4] Fu, W. J. (1998). Penalized regression: the Bridge versus the Lasso, *Journal of Computation and Graphical Statistics*, **7**, 397-416.

[5] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.

[6] Klinger, Artur. (2001). Inference in high dimensional generalized linear models based on soft thresholding. *Journal of the Royal Statistical Society B*, **63**, 377-392.

[7] Knight, K. and Fu, W. J. (2000). Asymptotics for Lasso-type estimators, *Annals of Statistics*, **28**, 1356-1378.

[8] Liu, Y., Zhang, H., Park, C. and Ahn, J. (2007). Support vector machines with adaptive L_q penalty. To appear in *Computational Statistics and Data Analysis.*

[9] Segal, M., Dahlquist, K. and Conklin, B. (2003). Regression approach for microarray data analysis. *Journal of Computational Biology*, 10, 961-980.

[10] Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989) Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: radical prostatectomy treated patients. *Journal of Urology*, **16**, 1076-1083.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society B*, **58**, 267-288.

[12] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net.*Journal of the Royal Statistical Society B*, **67**, 301-320.