PENALIZED PRINCIPAL COMPONENT REGRESSION

by

Ayanna Byrd

(Under the direction of Cheolwoo Park)

Abstract

When using linear regression problems, an unbiased estimate is produced by the Ordinary Least Squares. There are two serious drawbacks to using OLS; one is if $\boldsymbol{X}$ is less than full rank then the estimator is no longer unique and two if the design matrix is collinear then the estimates can have extremely large variances. To address these two problems, several penalized regression methods have been developed such as Ridge, Lasso and Bridge, all of which have improved OLS in some aspects. We study principal component (PC) regression with several different penalties to see if there is another way to improve the above-mentioned methods. Using various simulations and a real setting, we compare and contrast the different types of regressions methods to our PC regression methods. It is shown that PC regression in combination with two different penalties perform well in different situations when evaluating the simulation results .

Index words: penalized regression, Ridge, Lasso, Bridge, Principle Component

PENALIZED PRINCIPAL COMPONENT REGRESSION

by

Ayanna Byrd

B.S., Xavier University of LA, 2005

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

2008

PENALIZED PRINCIPAL COMPONENT REGRESSION

by

Ayanna Byrd

Approved:

Major Professor: Cheolwoo Park

Committee: Jeongyoun Ahn
William McCormick

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

Acknowledgments

I would like to thank my Family, especially my Mom and Dad for their love and support. They are the ones that instilled in me to believe in myself and to never give up during challenging periods.

My special thanks go to Dr. Lynne Seymour who has been there for me whenever I needed assistance and direction.

Thank you to my advisor, Dr. Cheolwoo Park, no words can express my gratitude toward you. His ability to guide me through and keep me on track is greatly appreciated.

I would like to thank my advisory committee members, Dr. Jeongyoun Ahn and Dr. William McCormick, who have provided valuable comments and directions for my research work for this thesis.

Finally, I would like to thank Ms. Xiong Yin who provided the initial R-code for penalized regression methods.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

When attempting to predict a real valued output y from a vector of inputs using standardized variables, consider the linear regression model

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

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 1.

The most used estimation method is the ordinary least squares (OLS) method, where we pick the vector of coefficients of $\hat{\boldsymbol{\beta}}$ to minimize the residual sum of squares (RSS).

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

which produces the best linear unbiased estimator (BLUE),

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

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

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

Although the OLS estimator is a BLUE, prediction accuracy and interpretation are two main problems that often occur when using OLS. The OLS estimator has the ability to have low bias, but large variance therefore causes one to have poor predictions and interpretations. By setting some of the coefficients to zero or shrinking them, we can possibly fix the problem

of poor prediction accuracy. After fitting a model, one would like to be able to explain the relationship between the predictor and response variables, where the predictor variables that are used in the model are the ones that explain the response variable the best.

There are several different shrinkage and variable selection methods that have been developed attempting to address the pitfalls of OLS. Several examples are as follows: Subset selection, Ridge regression, Lasso, and Principal Component (PC) regression.

Subset selection is a method of variable selection where the predicted model consists of a subset of variables and deletes the rest from the model. In subset selection, we look for the best subset. In order to estimate the coefficients of the subset of variables that are used, we use least squares regression. The size of the subset k is always less than or equal to the number of parameters ($k \leq p$), but how we determine k is another question. When determining k, we have to take into consideration the bias-variance tradeoff, where the chosen model is the model that minimizes the expected prediction error.

By choosing only a subset of predictor variables and deleting the rest, subset selection produces a model that is interpretable, but doesn't necessarily have a lower prediction error than the full model. Since the procedures only keep or throw away predictor variables, there is a tendency for the model to have a high variance, and for this reason other methods such as shrinkage methods are preferred.

Ridge regression (Hoerl and Kennard, 1970) is a shrinkage method where the assumption is made that the regression coefficients have a small possibility of being large after normalization. Ridge regression has a penalty on the size of the regression coefficients in order to shrink the coefficients. By adding the L_2 penalty term, we formulate the extension of OLS penalty term:

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

The ridge constant λ needs to be selected from data. The λ equal to zero corresponds to the least squares. As $\lambda \to \infty$, $\hat{\beta} \to 0$, meaning that the larger the λ the more the coefficients shrink. Ridge regression also helps to control high variance and minimize the estimate of

expected prediction error. Ridge regression is more stable than subset selection because it is a continuous process. A problem with ridge regression is that it doesn't give an easily interpretable model.

Lasso (Tibshirani, 1996), another shrinking method, is a combination of ridge regression and subset selection. The penalty and constraints of Lasso are slightly different than those of ridge regression, making some of the coefficients to have the value zero. Therefore, creating a model similar to subset selection where there is a subset of coefficients and the rest are zero (dropped). Similar to ridge regression and subset selection the constraint of lasso should minimize an estimate of expected prediction error. Lasso like Ridge is a continuous process, and more stable than subset regression, and it can also be applied to high-dimensional data. The Lasso method, however, has a problem with optimization when the number of predictor variables is larger than the number of observations. At most, the number of variables that are selected can be equal to the number of observations. Also, Lasso at times only selects one variable from pairwise correlated variables to help predict the response variable.

Bridge regression (Frank and Friedman,1993) is a method that reduces the residual sum of squares using the constraint $\sum |\beta_j|^q \leq t$ with $q \geq 0$. Having included the L_q penalty, the Bridge regression also includes the ridge regression and lasso when the q is equal to 2 and 1 respectively. When $0 \leq q \leq 1$, bridge regression performs subset selection. Subset selection and multicollinearity are easly addressed with Bridge regression because of the general form of the penalty. Bridge regression does have its drawbacks, one being that in order to perform bridge regression one has to solve the complex problem of nonconvex optimization, therefore it is not a widely used regression method.

PC regression (Massy, 1965) is another method where the goal is to choose a model with the lowest predicted MSE and desirable properties. The two properties that are used to determine whether or not to use the transformed model are (i) whether or not the variables are uncorrelated with each other and (ii) each of the predictor variables account for the variance of the full model. PC regression starts with the covariance matrix of the predictor

variables that are standardized.

$$\mathbf{V} = ave(\boldsymbol{x}\boldsymbol{x}^T) = \frac{1}{n}\boldsymbol{x}^T\boldsymbol{x}$$

and its eigenvector decomposition

$$\mathbf{V} = \sum_{k=1}^{p} e_k^2 \boldsymbol{v}_k \boldsymbol{v}_k^T,$$

where $\{e_k^2\}_1^p$ are the eigenvalues of V arranged in sequential order, with the largest value being first and $\{v_k\}_1^p$ are the corresponding eigenvectors. PC regression helps to deal with the problem of multicollinearity, where there are eigenvalues that are close to zero, so that the regression coefficients are unstable with large standard errors. A sequence of regression models is created by PCR with

$$\hat{\mathbf{y}}_\mathbf{k} = \sum_{k=0}^{K} [ave(y\boldsymbol{v}_k^T\boldsymbol{x})]\boldsymbol{v}_k^T\boldsymbol{x}, K = 1, ..., R.$$

Where R is the number of nonzero e_k^2. The main purpose of PC regression is to determine a particular model that $\hat{y}_k$ with the smallest mean square error. The reduced model has a set of variables that have successively smaller variances, where most of the variability of the original predictors is covered by the first principle component.

The thesis is organized as follows: in Chapter 2 we review the literature of Lasso and Bridge regression. In Chapter 3, we elaborate on PC regression with different penalties 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, fuel 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 and Bridge (L_q) estimator. All the methods are described using the standardized variables. That is,

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

2.1 Lasso

Least absolute shrinkage and selection operator (LASSO) (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. The Lasso estimate $\hat{\boldsymbol{\beta}}$ is defined by

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

or

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

Lasso has a penalty constraint called the $L_1 = \sum_j |\beta_j| \leq t$ which makes the solutions nonlinear in the y_i. Some of the coefficients become zero when t is extremely small, causing a type of subset selection to occur. If $t = t_0/2$, then the least squares coefficients are shrunk in half on average.

The Lasso translates each coefficient by a constant factor, truncating at zero, called 'soft-thresholding'. Where the thresholding parameter λ in the Lasso formula is a one-to-one

transformation of the bound t appearing in the definition. 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 Bridge

The Bridge regression coefficients can be obtained by :

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

Fu (1998) sugessted a comptutational alogrithm for $q \geq 1$, when dealing with Bridge regression (or L_q regression), and Xiong (2007) suggested an algorithm for $q > 0$. Letting the q be estimated from the given data, the L_q extends the practicality of the Bridge estimator. When noise and redundant variables exist, the Bridge estimator needs variable selection, and the q are estimated to be less than or equal to one. In other cases, the q will be greater than 1. Note that Lasso ($q = 1$) and Ridge ($q = 2$) are special cases of L_q regression. Xiong (2007) found that the L_q is very robust and performs better than other well known regression procedures such as the OLS, Ridge, and the Lasso.

In practice, a learning procedure of L_q penalty with a fixed q has its advantages only under certain situations because different penalties may perform better for different data structures; therefore, q is treated as a tuning parameter. Fan and Li (2001) denote a loss function of $l(\beta)$ and proposed the following algorithm to minimize penalized general loss by local quadratic approximations $l(\beta) + n \sum_j p_\lambda(|\beta_j|)$ where p_λ is a penalty function. 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 $p'_\lambda(|\beta_j|) = \lambda q|\beta_j|^{q-1}$. The solution for the penalized least squares problem can be found by iteratively computing the Ridge regression (Fan and Li, 2001 or Xiong, 2007).

For L_q regression, there are two tuning parameters λ and q. The parameter λ controls the trade-off between minimizing the loss and the penalty. The other tuning parameter q determines the penalty function, where it is important to choose the correct q. If there is a large number of noise input variables, the L_q penalty with $q \leq 1$ is desired. If all the covariances are important, it might be better to use $q > 1$ to avoid unnecessary variable deletion. L_q regression finds the optimal combination of λ and q, where the penalty function is minimized.

Chapter 3

PC Regression

3.1 Introduction

Principal Component (PC) regression is a method used to estimate response variables based on principal components of the explanatory variables. PC regression is a regression method that addresses the problem of multicollinearity. The main reason why one would want to regress the response variables on the principal components of the explanatory variables is because of the multicollinearity. Highly correlated explanatory variables (multicollinearity) may cause unstable estimations of the least squares coefficient. Multicollinearity happens when one or more eignevalues of $(X'X)$ are close to zero, so that $(X'X)^{-1}$ becomes unstable. Thus the corresponding linear regression model will be

$$y_j = \beta_0 + \beta_1 PC_{1j} + ... + \beta_M PC_{Mj} + \epsilon_j$$

where $PC_{1j}, ..., PC_{Mj}$, $j = 1, ..., n$ are respectively the principal component scores of the PC's with $M < p$. Although we are only using the M PC's, all p original predictor variables are still represented in the model through them. If $M = p$, then the regression is equivalent to the ordinary least squares.

In this thesis, we take the matrix of the predictor variables V=ave(XX') and its eigenvector decomposition

$$\mathbf{V} = \sum (e_p^2 v_k v_k^T),$$

where $\{e_k^2\}_1^p$ are eigenvalues of V arranged in a decreasing order and $\{v_k\}_1^p$ are their corresponding eigenvectors. These eigenvectors that we get from the decomposition are called principal components, where principal component 1 corresponds to eigenvector 1. The number

of variables determines the number of principal components that can be constructed, each of which is a linear combination of the original set of variable. If there are p-variables, then we are able to construct p-principal components assuming that $p < n$.

Instead of using the cumulative percentage cutoff to determine which principal components are best suited to use in order to have a reduced mean square error, we will use the lasso technique and the Akaike Information Criteria (AIC) as a penalty to see if we have a better model. The AIC criterion is commonly used model selection technique that penalizes for the complicated model. In general,

$$\mathbf{AIC_M} = -\frac{2}{n} loglikelihood + 2\frac{M}{n}$$

where M is the number of parameters in the model. When selecting the model (determining M) we want to minimize the AIC so we will take the first M principal components that we find and use them as our x values and our original y's to create our model. Hopefully once this is done we will have a model that has a lower MSE than when just using OLS, Lasso, Ridge, and L_q individually.

3.2 Computation

In order to formulate the two suggested methods of regression we used the following algorithms. When formulating the method for PC+Lasso our first step is to find the PC's of the explanatory variables X. Once we have our principal components using eigenvalue decomposition and standardize the principal components, we use the $\boldsymbol{Z}$ to represent our new explanatory variable matrix and apply the Lasso method. After completing this procedure we estimate $\hat{\boldsymbol{\beta}}$ by

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

or

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

When formulating the method for PC+AIC we again use the matrix $\boldsymbol{Z}$. We estimate the PC+AIC estimate by:

$$\hat{\boldsymbol{\beta}}_{PC+AIC} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{m=1}^{p} \hat{\beta}_m Z_{mi})^2 + \frac{2M}{n} \right\}.$$

To obtain $\hat{\beta}_{PC+AIC}$, we do the following steps:

- Step 1: Set $m = 1$ and $AIC_0 =$ large value.
- Step 2: Find $\hat{\beta}_{OLS}^{(m)}$ using the first m PCs.
- Step 3: Calculate AIC_m using the estimated $\hat{\beta}_{OLS}^{(m)}$.
- Step 4: If $AIC_m < AIC_{m-1}$, $m \leftarrow m + 1$ and go back to Step 2. If not, $M = m - 1$ and $\hat{\beta}_{PC+AIC} = \hat{\beta}_{OLS}^{(M)}$.

Chapter 4

Simulation

In order to find out how good our PC estimators are in comparison to other penalized regression techniques, we performed a simulation study using OLS, Ridge regression, Lasso, L_q, PC+Lasso, and PC+AIC.

4.1 Simulation Settings

There are five different settings in the simulation. We simulate data from the model $\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\sigma}\boldsymbol{\epsilon}$, ϵ $N(0, 1)$. The first three settings, which were used in the original Lasso paper (Tibshirani, 1996), main purposes are to systematically compare the prediction performance of Lasso and Ridge regression. In the fourth setting the main purpose is to compare how the different techniques address high collinearity. The fifth 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.

Three data sets are used in each simulation: 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 $\sigma = 3$ has a multidimensional normal distribution with mean 0. 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:

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

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

4. Setting 4: The same as setting 3 except that the $corr(i, j) = 0.8$ for all i and j.

5. Setting 5: 50 simulated data sets, each consisting of 50/50/400 observations and 40 predictors. The predictors were set to be:

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

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.

As a unit of measure for the simulation results, we used the mean square error. 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	L_q	PC+Lasso	PC+AIC (M)
1	Mean	0.236	0.200	0.188	0.185	0.247	0.359 (2.280)
	s.e	0.026	0.016	0.016	0.016	0.018	0.027 (0.164)
	Median	0.209	0.209	0.167	0.162	0.237	0.333 (2.000)
2	Mean	0.716	0.213	0.291	0.233	0.214	0.145 (1.740)
	s.e	0.040	0.014	0.023	0.020	0.021	0.016 (0.117)
	Median	0.634	0.179	0.242	0.176	0.174	0.174 (2.000)
3	Mean	0.156	0.050	0.065	0.050	0.123	0.039 (1.220)
	s.e	0.006	0.002	0.002	0.002	0.009	0.001 (0.066)
	Median	0.155	0.049	0.066	0.047	0.118	0.039 (1.000)
4	Mean	0.105	0.018	0.031	0.019	0.056	0.014 (1.200)
	s.e	0.004	0.001	0.001	0.008	0.002	0.001 (0.069)
	Median	0.103	0.018	0.030	0.018	0.052	0.012 (1.000)
5	Mean	1.230	0.145	0.141	0.111	0.521	0.160 (2.820)
	s.e	0.086	0.006	0.001	0.021	0.009	0.039 (0.184)
	Median	1.087	0.148	0.133	0.064	0.519	0.066 (3.000)

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

In all five 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: $(0.1, 0.4, 0.7, 1, 1.3, 1.7, 2, 2.5, 3)$.

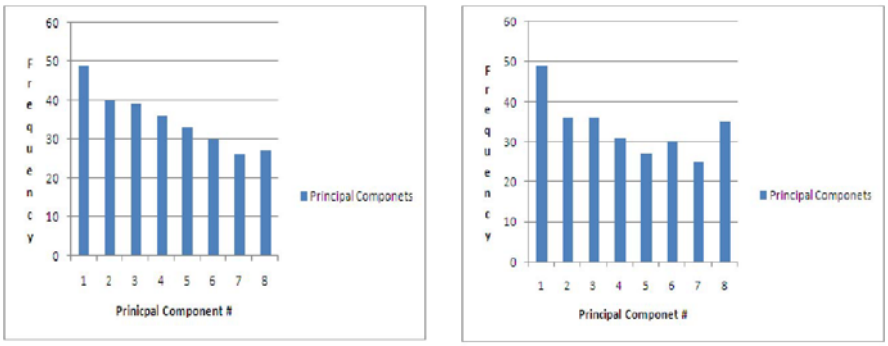
4.2 Simulation Results

Table 4.1 shows the results of our simulation. The PC+Lasso does better than least squares, but is about the same as the Ridge, Lasso, and L_q except in simulation 5. PC+AIC is the best

in several situations. The first simulation setting has 8 predictor variables and the true values of 5 out of 8 are 0 with a weak correlation. Based on previous simulations (Xiong, 2007), we would expect Lasso and L_q to have similar results. When looking at the results, we see that indeed this happens with Lasso having a mean mse of .188 and L_q a mean mse of .185. None of the PC combinations, do any better than an already established regression method. We see in the Figure 4.1(a) the number of coefficients in PC+Lasso declines as expected, but there is not a harsh drop where some of the the coefficients are zero. In PC+Lasso all of the PC's tend to be selected in order from PC 1 to PC 2 and the first PC is selected in all but a few cases. When using PC+AIC, $M = 2$, meaning that the first two PC's where commonly selected. In both, the PC+AIC and the PC+Lasso, this is somewhat expected because the early PC's mainly the first PC is designed to explain most of the data.

The second simulation setting also has 8 predictor variables; however, this time the true values of all the 8 β's are 0.85, and there is a weak correlation, which means no variable selection is intended. Some correlation structures exist; therefore, Ridge regression ($q = 2$) is expected to perform the best. In fact, Ridge regression performs second best to PC+AIC. PC+Lasso performed very similar to Ridge regression. The mean mse for PC+AIC is 0.145, where the mean mse for Ridge is 0.213, and for PC+Lasso is 0.214. We see in the Figure 4.1(b) like in the first simulation the number of coefficients in PC+Lasso declines as expected, but there is not a harsh drop where some of the coefficients are zero. As in the first simulation, in PC+Lasso all of the PC's tend to be selected in order from PC 1 to PC 2 and the first PC is selected in all but a few cases. When using PC+AIC, $M = 2$ meaning that the first two PC's where commonly selected. In both the PC+AIC and the PC+Lasso, this is again is somewhat expected because of the nature of principal components.

The third simulation setting has 40 predictor variables, and there is high collinearity among these variables ($\rho = .5$). Also there are 20 zeros out of the 40 coefficients and the noise level was increased to $\sigma = 15$. Again, PC+AIC performs the best with a mean mse of 0.039. Ridge follows with a mean mse of .05 while PC+Lasso only does better than the



(a) Simulation 1 (b) Simultion 2

Figure 4.1: Frequency of nonzero βs of PC+LASSO in 50 repetitions (simulations 1 and 2)

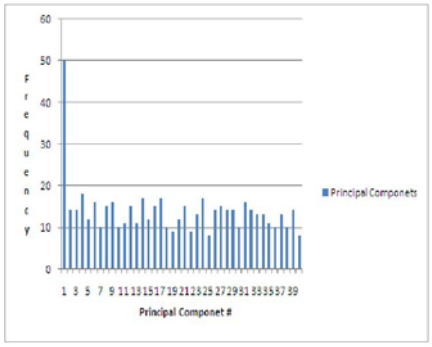


Figure 4.2: Frequency of nonzero βs of PC+LASSO in 50 repetitions (simulation 3)

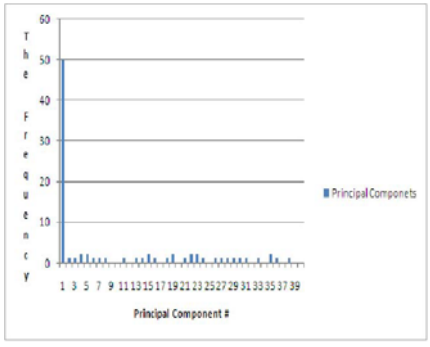


Figure 4.3: Frequency of nonzero βs of PC+LASSO in 50 repetitions (simulation 4)

OLS with a mse of .123. We see in the Figure 4.2 the number of coefficients in PC+Lasso declines as expected, but unlike in the first two simulations we do see the effects of the harsh-thresholding that occur from Lasso. In PC+Lasso all of the PC's tend to be selected in no real order though the first PC is selected in every case. There is even more of a drop off between the number of the first PC and the number of the other 39 PC in comparison to the first and second simulation, but once there is a drop off the number of a particular PC used increases and decreases. When using PC+AIC, $M = 1$ meaning that the first PC was the commonly selected. In both the PC+AIC and PC+Lasso, the number of PC's before the harsh drop off is similar to the small number of PC's used in PC+AIC. The drop off and the small number of PC's are expected because there are many noisy variables.

Like in third simulation setting, the forth simulation setting has 40 predictor variables, and there is high collinearity among these variables. However, this time the collinearity is even higher ($\rho = .8$). PC+AIC does the best with a mean mse of .014. This time PC+Lasso does not act so much like the ridge regression as in previous simulations; in fact, it acts worse, but better than lasso with a mse mean of .026. Like in the third similation, we see in the Figure 4.3 the number of coefficients in PC+Lasso declines as expected; unlike in the first two simulations, we do see the effects of the harsh-thresholding that occur from Lasso. In this case, we do see many of the number of coefficients equal zero. In PC+Lasso, all of the

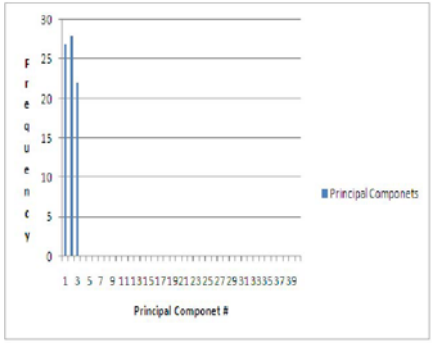


Figure 4.4: Frequency of nonzero βs of PC+LASSO in 50 repetitions (simulation 5)

PC's tend to be selected in no real order, with the first PC being selected every time. There is even more of a drop of between the number of the first PC and the number of the other 39 PC in comparison to the first and second simulation, but once there is a drop off the number of time a particular PC is selected no more than five times. When using PC+AIC, $M = 1$ meaning that the first PC was the commonly selected.

The fifth simulation setting deals with group effect. Also, 15 out of 40 coefficients are zeros. Both of the PC regressions out perform the OLS. PC+Lasso has a mean mse of .521, while PC+AIC performs similarly to Ridge, Lasso, and L_q with a mean mse of .1603. We see in the Figure 4.4 the number of coefficients in PC+Lasso declines as expected, but unlike the in the first two simulations we do see effects of the harsh-thresholding that occur from Lasso most of the coefficients go to zero. In PC+Lasso all of the PC's tend to be selected in order. The first PC is now joined by the second and third PC having been selected in most of the case, but this time the second PC is actually selected more often. There are even more of a drop of between the number of the first three PC's and the rest of the PC's, but once there is a drop off the number of a particular PC that is selected is zero across the board, showing the harshest drop off of all five settings. When using PC+AIC, $M = 3$ meaning that the first three PC's where commonly selected. These three PC's seem to correspond to the three groups in the simulation setting.

In summary, the PC+AIC is the best option in three out of the five settings and is always better the OLS in four out of the five. This shows that PC+AIC is very robust, outperforming at least one of the other regression methods. Even though the number of the PC's that are used in PC+AIC and the number of PC's used before the threshholding occurs in the PC+Lasso are about the same, the two methods PC+Lasso and PC+AIC perform differently. PC+Lasso at times performs similar to ridge regression and better than at least OLS, and other times it even performs better than PC+AIC, but it never performs the best overall.

Chapter 5

Fuel Data Analysis

The fuel data come from a study by Weisberg (1985) that examined the correlation between motor fuel consumption and a number of variables recorded that are related to the ability to drive a vehicle. The study happened over a period of two years from 1971-72, where there are a total of 48 observations which represent the 48 contiguous states. Of the 48 data points, 24 were used for the training set and 24 were used for testing. The response variable is $fuel$ - the motor fuel consumption in gallons per person. The explanatory variables are as follows:

1. pop: 1971 Population, in thousands
2. tax: 1972 Motor fuel tax rate, in cents per gallon
3. nlic: 1971 Thousands of licensed drivers
4. inc: 1972 Per capita income in thousands of dollars
5. road: 1971 Thousand of miles of federal-aid primary highways
6. fuelc: 1972 Fuel consumption, in millions of gallons
7. dlic: 1971 Percentage of population with driver‘s license

Figure 5.1 shows that five of the variables have a linear relationship with fuel, where dlic has a positive trend and pop, tax, nlic and fuelc seem to have a negative trend. There seems to be no relationship between fuel and inc or road. Several of the variables are related with each other for example, there seems to be a positive relationship between pop vs. nlic, pop vs. fuelc, nlic vs. fuelc, fuelc vs. road.

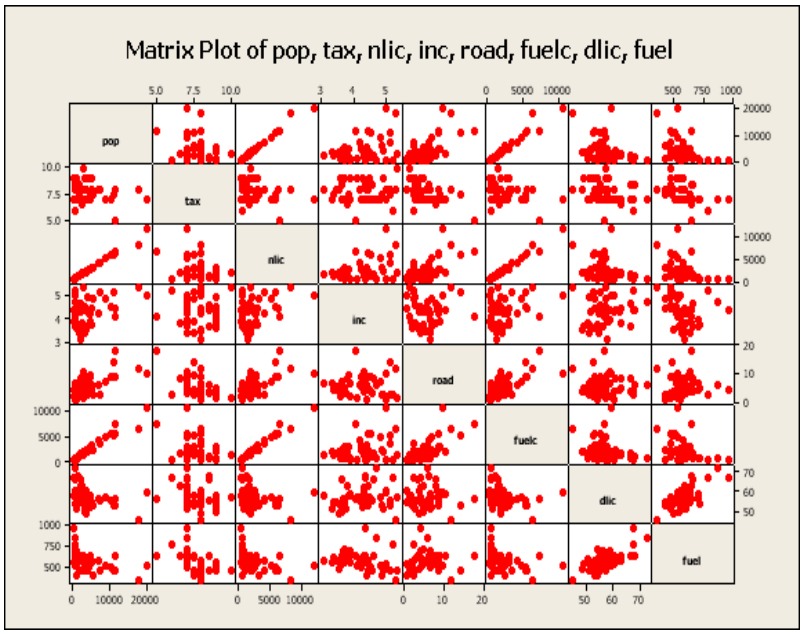


Figure 5.1: Scatter diagram matrix.

Correlations: pop, tax, nlic, inc, road, fuelc, dlic, fuel

	pop	tax	nlic	inc	road	fuelc	dlic
tax	-0.147						
	0.320						
nlic	0.991	-0.180					
	0.000	0.222					
inc	0.410	0.013	0.404				
	0.004	0.932	0.004				
road	0.646	-0.522	0.648	0.050			
	0.000	0.000	0.000	0.735			
fuelc	0.970	-0.237	0.988	0.333	0.702		
	0.000	0.105	0.000	0.021	0.000		
dlic	-0.365	-0.288	-0.299	0.157	-0.064	-0.286	
	0.011	0.047	0.039	0.286	0.665	0.048	
fuel	-0.463	-0.451	-0.423	-0.245	0.019	-0.363	0.699
	0.001	0.001	0.003	0.093	0.898	0.011	0.000

Cell Contents: Pearson correlation
P-Value

In the correlation matrix we see that several variables have a strong correlation with fuel. Fuel has a very strong positive correlation with dlic ($r = 0.699$, $P = 0.000$), and significant negative correlation with pop ($r = -0.463$, $P = 0.001$), tax ($r = -0.451$, $P = 0.001$), nlic ($r = -0.423$, $P = 0.003$) and fuelc ($r = -0.363$, $P = 0.011$). There are also several explanatory variables that are strongly correlated with one another. For example, there is a significant positive correlation between pop vs. nlic ($r = 0.991$, $P = 0.000$) and pop vs. fuelc ($r = 0.988$, $P = 0.000$). Road is also has positive correlation with several of the other variables, pop ($r = 0.646$, $P = 0.000$), nlic ($r = 0.648$, $P = 0.000$), and fuelc ($r = 0.702$, $P = 0.000$).

Table 5.1: Estimated Model Errors for Fuel Data

Sim	MSE	OLS	Ridge	Lasso	L_q	PC+Lasso	PC+AIC (M)
1	Mean	0.500	0.453	0.473	0.471	0.453	0.442 (2)

With this highly correlated data, our simulation results are supported and the different regression methods perform as we expected. Table 5.1 shows us that again, PC+AIC performs the best with the smallest mse of .442 where 2 principal components where used as predictor variables. Performing similarly PC+Lasso and ridge follow with PC+Lasso doing slightly better than ridge with a mse of .453 and .457 respectively. PC+Lasso used all of the principal components except number 5. L_q, and Lasso also perform similarly with mse of .471 and .473 respectively, where L_q used all six variables except road and Lasso used all seven explanatory variables. As expected, OLS does the worst with a mse of .500

Chapter 6

Summary

In this thesis, we studied PC regression along with added penalties. We have presented some evidence that suggesting that the PC+Lasso and PC+AIC perform similarly if not better than OLS, Ridge, Lasso, L_q regression methods. We examined the relative merits of the methods in five different scenarios:

1. variable selection – Lasso and L_q performs the best closely followed by Ridge, OLS, and PC+Lasso, PC+AIC brings up the rear;

2. correlation beween variables – PC+AIC performs the best, Ridge and PC+Lasso closely follow;

3. high collinearity and variable selection – PC+AIC perform the best, Ridge, Lasso and L_q closely follow, PC+Lasso performs better than OLS;

4. extremely high collinearity and variable selection – PC+AIC performs the best, Ridge,Laso and L_q closely follow, PC+Lasso performs better than OLS;

5. group effect and variable selection – L_q performs the best, Ridge, Lasso and PC+AIC closely follow, with PC+Lasso performing better than OLS.

The two types of PC regression is a method where we use the principal components in combination with the Lasso corresponding to $q = 1$ as a constraint or the minimized AIC as a constraint. The simulation result shows that PC+AIC regression is a very robust regression method outperforming the other methods three out of the five methods. When there is high collinearity with a correlation of $\rho \geq .5$, one can use PC+AIC because it out performs the

all of the methods. PC+Lasso performing close to Ridge, Lasso and L_q in several situations shows that the one could choose PC+Lasso and it has the possibility of out performing either three. The results suggest that both the PC+Lasso and the PC+AIC might be very useful when attempting to solve a problem of statistical estimation. One might be more inclined to choose PC+AIC more often than PC+Lasso, but at times one might choose PC+Lasso. Further investigation is needed to truly understand and support these findings.

Bibliography

[1] 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.

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

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

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

[5] Johnson, Richard Arnold. (1998). Applied multivariate statistical analyisis,*Prentice Hall*, 458-498.

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

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

[8] Weisberg, S. (1985). Applied Linear Regression, Wiley.

[9] Xiong, Yin (2007). L_q Penalized Regression, *Masters Thesis*, University of Georgia.