Scientific Research Publishing

Advances in Pure Mathematics, 2017, 7, 51-65
http://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Group Variable Selection via a Combination of L_q Norm and Correlation-Based Penalty

Ning Mao, Wanzhou Ye

Department of Mathematics, College of Science, Shanghai University, Shanghai, China
Email: wzhy@shu.edu.com

How to cite this paper: Mao, N. and Ye, W.Z. (2017) Group Variable Selection via a Combination of L_q Norm and Correlation-Based Penalty. *Advances in Pure Mathematics*, **7**, 51-65.
http://dx.doi.org/10.4236/apm.2017.71005

Received: December 16, 2016
Accepted: January 21, 2017
Published: January 24, 2017

Abstract

Considering the problem of feature selection in linear regression model, a new method called LqCP is proposed simultaneously to select variables and favor a grouping effect, where strongly correlated predictors tend to be in or out of the model together. LqCP is based on penalized least squares with a penalty function that combines the L_q $(0<q<1)$ norm and correlation-based penalty that is CP norm. It can shrink some coefficients to exactly zero and additionally the CP term links strength of penalization to the correlation among the predictors. The simulation studies show the advantages of LqCP with the increase of noise variables and the case of $p>n$. In addition, a simulation about grouped variable selection is performed. Finally, The model is applied to two real data: US Crime Data and Gasoline Data. In terms of prediction error and estimation error, empirical studies show the efficiency of LqCP.

Keywords

Linear Regression, Variable Selection, Elastic Net, Adaptive Elastic Net

1. Introduction

Here the usual linear regression mode is considered in the paper given by:

$$y = X\beta + \epsilon, \tag{1}$$

where $y_{n\times 1}$ are the observations, $\beta_{p\times 1}$ is a vector of unknown parameters to be estimated, $X_{n\times p}$ is an $n\times p$ matrix of p predictor vectors of n observations and ϵ is a random error vector with $E(\epsilon)=0$ which often is assumed that it is subject to normal distribution independently. Ordinary least squares (OLS) estimates are very common which can be obtained by minimizing the sum of square residuals. In general, the criteria for evaluating the quality of a

model from the following two aspects. One is prediction accuracy on test data and the other is to tend to select a simple model. In other words, less variables would be selected in the condition of same prediction effects. Variable selection is necessary especially when the number of predictors is large. There are many applications using variable selection to solve problems like knowledge discovery with high-dimensional data source [1] and it could greatly enhance the prediction performance of the fitted model. Traditional model selection method is best-subset selection and its step-wise variants. However, best-subset selection is computationally prohibitive when the number of predictors is large. As analyzed by Breiman (1996) [2], subset selection is unstable; thus, the resulting model has poor prediction accuracy. To overcome the drawbacks of subset selection, many variable selection methods are appeared and the most popular recently is regulation method.

In recent years, regularization method has attracted a great attention. It is used in applications such as machine learning, denoising, inpainting, deblurring, compressed sensing, source separation and more. Generally, the square loss function of penalized least squares estimates:

$$L_{(\beta;\lambda)} = \|y - X\beta\|_2^2 + \lambda\|\beta\|_q^q, \tag{2}$$

where $\lambda > 0$ is a penalty parameter. $q > 0$ is considered in this paper. $\|\beta\|_q^q = \sum_{j=1}^{p}|\beta_j|^q$. When $q = 1$, it becomes Lasso procedure. The procedure of minimizing the objective function is called ridge regression when $q = 2$. As a continuous shrinkage method, ridge regression achieves its better prediction performance through a bias-variance trade off. However, ridge regression can not produce parsimonious model, which means all the predictors are kept in the model. Lasso is proposed by [3] in 1996. It imposes the L_1 norm regularization to the loss function and becomes a widely popular regularization method. Further, Knignt and Fu (2000) [4] studied the asymptotic properties of lasso. Lasso can shrink some coefficients to zero to achieve the effects of variable selection. Due to this reason, lasso has gained popularity in high-dimensional data.

Although lasso is a popular method for variable selection, it still has several drawbacks. The first is the lack of oracle property. The oracle property means the probability of selecting the right set of nonzero coefficients converges to one, and the estimators of the nonzero coefficients have asymptotically normal distribution with the same means and covariances. Fan and Li (2001) [5] first pointed the parameters estimators of lasso are biased and don't have the oracle property. Then adaptive lasso proposed by Zou (2006) [6] overcomes this limitation of lasso. In adaptive lasso, weights are used for regularizing different coefficients in the L_1 norm regularization, which means different coefficients have different shrinkage factors. Secondly, lasso have very poor performance when there are highly correlated variables in high-dimensional data. What's more, lasso only selects n variables at most when the number of variables p is more than the number of sample observation size n, which is a big limitation

in application areas. At last, there are grouping variables among genes for the microarray gene data. These genes mostly have a high pairwise correlation with each other while lasso can not select a group of correlated genes. Based on these limitations, Zou and Hastie [7] proposed the elastic net regularization which is a combination of L_1 norm and L_2 norm. Similarly, elastic net (ENET) lacks the oracle property even though it outperforms Lasso. Further, Zou and Zhang [8] proposed adaptive elastic net to achieve the oracle property and good estimation accuracy. In the base of that, Ghosh [9] further studied grouping effects in the adaptive elastic net by using the ordinary least squares as the initial weight in low dimension data for simplicity. However, the estimators of ordinary least squares is very bad in the case of large numbers predictors or high correlated variables. Additionally, elastic net and adaptive elastic net don't take into account the information about the correlations of variables.

In the same spirits, there existed other penalty based on methods for handling grouping effects. Penalizing least squares via combining L_1 and L_∞ named OSCAR is presented by Bondell and Reich (2008) [10]. The Oscar forces some coefficients to be identically equal, encouraging correlated variables that have similar effects on the response to form clusters represented by the same coefficients. However, the computation of Oscar estimation is slow for large p which is based on a sequential quadratic programming. Then, considering the information of correlations of variables, Tutz and Ulbricht (2009) researched the property of correlation-based penalty (CP) and pointed the grouping effect of CP term. In the article of Tutz and Ulbricht, blockwise boosting procedure is applied in the simulations, which updates at each step the coefficient of more than one variable. However, in practical implementation, the step length factor and the stopping number of iterations have to be determined. Sometimes, this may be difficult and affects the sparsity of the solution as well as the speed of convergence of the algorithm. Therefore, El Anbari and Mkhadri (2014) [11] proposed an alternative regularization procedure called L1CP by combing L_1 norm with CP term. The method performs automatic variable selection and has the ability of grouping effects.

In this paper, motivated by the sparsity and grouping effect especially the case of the pairwise correlations are very high, a new regulation procedure called LqCP is proposed in linear regression setting. It combines the L_q $(0 < q < 1)$ norm and CP penalty. Similar to the L1CP method, LqCP also performs automatic variables selection and allows to select or to remove highly correlated variables together. Section 2 first introduces the property of elastic net and adaptive elastic net and then demonstrates LqCP and its algorithm as well as the method of choosing regularization parameters. In Section 3, simulation studies give the estimation mean squared errors for different circumstances to show the parameters estimation effects among LqCP, elastic net (ENET), adaptive elastic net (AENET) and L1CP. Section 4 is mainly about the application examples to show the prediction accuracy of models. The conclusion of this paper is given in Section 5.

2. Methods

2.1. The Elastic Net and Adaptive Elastic Net

Here the form of elastic net described in the above firstly is showed in the following. The naive elastic net estimator $\hat{\beta}_{(\text{naive})}$ is the minimizer of equation:

$$\hat{\beta}_{(\text{naive})} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\| + \lambda_2 \|\beta\|_2^2 . \quad (3)$$

This method is called the naive elastic net which overcomes the limitations of Lasso in the case of $p > n$. The penalty of combining L_1 norm and L_2 norm which has been proved that it has grouping effect in Zou and Hastie (2005) [7]. The estimator of elastic net is a two-stage procedure: for each fixed λ_2, they first find the ridge coefficients, and then do the Lasso-type shrinkage along the lasso coefficient solution paths. At last, through the criterion of model selection, the best λ_1 and λ_2 are obtained. It appears to incur a double amount of shrinkage. Double shrinkage does not help to reduce variances much and introduces unnecessary extra bias, compared with pure lasso or ridge shrinkage. Hence, the elastic net estimators are rescaled by $\hat{\beta}_{(\text{enet})} = (1+\lambda_2)\hat{\beta}_{(\text{naive})}$. Such a scaling transformation preserves the variable selection property of the naive elastic and improves the prediction performance by correcting this double shrinkage.

In a similar way to Lasso, the elastic net does not enjoy the oracle property. Combining the property of the adaptive Lasso, Zou and Zhang [8] proposed the adaptive elastic net which combines the adaptive Lasso and L_2 norm. Additionally, they established the oracle property of the adaptive elastic net in the condition of weak regularity. The form of the adaptive elastic net is as follows

$$\hat{\beta}_{(\text{Adaenet})} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} \hat{\omega}_j |\beta_j| + \lambda_2 \|\beta\|_2^2 , \quad (4)$$

where $\hat{\omega}_j = \left(|\hat{\beta}_j|\right)^{-\gamma}$, $j = 1, 2, \cdots, p$, and γ is a positive constant. Choosing the initial weight is crucial in the adaptive elastic net. Zou and Zhang [8] proposed using the elastic net as an initial weight either in low-dimensional data or high-dimensional data. In the paper of Ghosh [9], the weights is donated by $\hat{\beta}_{(ols)}$. Ghosh [9] researched the ability of variable selection and grouping effect problem, proved that the adaptive elastic net also can automatically select the group variables and also had a better performance for prediction accuracy than elastic net. But the estimator of ordinary least squares is very bad in the case of high dimension or high correlations of variables. Therefore, based on the paper of Zou [8], the elastic net estimators is regarded as the initial weight in simulations in this paper. To avoid the invalid values when the estimators $\hat{\beta}_j = 0$, the initial weights are defined by $\hat{\omega}_j = \left(|\hat{\beta}_j| + 1/n\right)^{-\gamma}$, where n is the sample size and $\gamma > 0$. For adaptive methods, while as $\gamma \to 0$ it becomes a biased procedure like lasso. Similarly for $\gamma \to \infty$ adaptive methods are unbiased. A good choice of γ is a compromise between shrinkage and bias and can be obtained via cross-validation. As described in Ghosh, γ should not be

chosen too big a quantity and typically first should be given a range of values. Then for each fixed γ, the cross-validation is applied to select other tuning parameters. But the elastic net and adaptive elastic net does not take into account the correlations structure of variables. Therefore, according to the CP term proposed by Tutz and Ulbricht (2009) [12], proposed new method is as follows.

2.2. The Proposed Method

2.2.1. Introduction of Proposed Model

In the context of linear regression problems, the following penalty function based on residual sum squares is considered. The L_q penalty on f is defined as

$$\|f\|_q^q = \sum_{j=1}^{p} |\beta_j|^q. \tag{5}$$

When $q=0$, the corresponding penalty is discontinuous at the origin and consequently is not easy to compute. Thus in this paper $q>0$ is designed. The least squares subject to the L_q penalty with $0<q<1$ is first studied by Frank and Friedman (1993) [13] which is known as bridge regression. Fu (1998) [14] and Knight and Fu (2000) [4] studied asymptotic properties and the computation of bridge estimators. When $q=2$, the solution $\hat{\beta}$ never becomes zero unless $\hat{\beta}=0$ and it is biased. For $q\le 1$, the bridge estimator tends to shrink small absolute coefficients to exact zeros and hence selects important variables. As pointed out by Theorem 2 in Knight and Fu (2000) [14], when $q>1$ the amount of shrinkage towards zero increases with the magnitude of the regression coefficients being estimated. It suggests that if $q<1$, estimate nonzero regression parameters at the usual rate without asymptotic bias while shrinking the estimates of zero regression parameters to 0 with positive probability. In practice, in order to avoid unacceptable large bias for large parameters, the value of q is often chosen not too large. When $0<q<1$, the L_q penalty may achieve better sparsity than L_1 penalty because larger penalty is imposed on small coefficients than L_1 penalty.

So according to better sparsity property of L_q $(0<q<1)$ and the case of high correlations of variables, the model is defined by

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_q^q + \lambda_2 P_c(\beta), \tag{6}$$

where

$$P_c(\beta) = \sum_{j=1}^{p}\sum_{i>j} \frac{(\beta_i - \beta_j)^2}{1-\rho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1+\rho_{ij}}. \tag{7}$$

$0<q\le 1$, λ_1, λ_2 are positive constants. ρ_{ij} denotes the (empirical) correlation between the ith and jth predictors. Here the penalty $P_c(\beta)$ is introduced by Tutz and Ulbricht (2009) [12] and is defined assuming that $\rho_{ij}\neq 1$ for $i\neq 1$. When $q=1$, the model changes to the combination of L_1 and $P_c(\beta)$, which is the model called L1CP proposed by El Anbari M. and Mkhadri A. (2014) [11].

As introduced in the above, L_q penalty in the loss of residual square sum have a better sparsity, while the correlation-based penalty $P_c(\beta)$ will encourage grouping effect for highly correlated variables. In fact, it's easy to see that for strong positive correlation ($\rho_{ij} \approx 1$) in (7), the first term becomes the dominant having the effect that estimates for β_i and β_j are similar ($\hat{\beta}_i \approx \hat{\beta}_j$). For strong negative correlation ($\rho_{ij} \approx -1$), the second term becomes dominant and β_i will be close to β_j. The effect is grouping, highly correlated effects show comparable values of estimates ($\left|\hat{\beta}_i\right| \approx \left|\hat{\beta}_j\right|$) with the sign being determined by positive or negative correlation.

Moreover, assuming that $\rho_{ij} \neq 1$ for $i \neq j$, the penalty (7) can be written in a sample quadratic form $P_c(\beta) = W^{\mathrm{T}} \beta W$, where $W = \left(w_{ij}\right), 1 \leq i, j \leq p$, is a positive definite matrix with general term

$$W_{ij} = \begin{cases} 2\sum_{s \neq i} \dfrac{1}{1-\rho_{is}^2} & \text{if } i = j \\ -2\dfrac{\rho_{ij}}{1-\rho_{ij}^2} & \text{if } i \neq j \end{cases} \tag{8}$$

Putting $\gamma = \lambda_2 / (\lambda_1 + \lambda_2)$,the optimization problem (6) is equivalent to

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2, \text{ s.t } (1-\gamma)\|\beta\|_q^q + \gamma P_c(\beta) \leq \nu \text{ for } \nu \geq 0. \tag{9}$$

Proof of (8): In fact, the penalty $Q(\beta) = \lambda_1 \|\beta\|_q^q + \lambda_2 P_c(\beta)$ can be written as follows:

$$\begin{aligned} Q(\beta) &= \lambda_1 \|\beta\|_q^q + \lambda_2 P_c(\beta) \\ &= (\lambda_1 + \lambda_2)\left\{\frac{\lambda_1}{\lambda_1 + \lambda_2}\|\beta\|_q^q + \frac{\lambda_2}{\lambda_1 + \lambda_2} P_c(\beta)\right\} \\ &= \lambda\left\{(1+\gamma)\|\beta\|_q^q + \gamma P_c(\beta)\right\} \end{aligned} \tag{10}$$

where $\lambda = \lambda_1 + \lambda_2$. So, the problem (6) is equivalent to finding

$$\hat{\beta}(\lambda, \gamma) = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\left\{(1-\gamma)\|\beta\|_q^q + \gamma P_c(\beta)\right\} \tag{11}$$

which is equivalent to the optimization problem (6).

2.2.2. The Algorithm Procedure

The estimators of β can be computed via the Cyclic Descent Algorithm for l_q sparsity penalized linear regression problem [15]. The main idea is to transform the LqCP problem into an equivalent problem on augmented data.

Indeed, the optimization problem (6) can be written as

$$\hat{\beta}(\lambda_1, \lambda_2) = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_q^q + \lambda_2 \beta^{\mathrm{T}} W \beta, \tag{12}$$

where W is defined by (8), is a real symmetric positive-define square matrix, assuming that $\rho_{ij}^2 \neq 1$, with Choleski decomposition $W = LL^{\mathrm{T}}$ and $L = W^{\frac{1}{2}}$, which always exists. Now, let

$$X^*_{(n+p)\times p} = \begin{pmatrix} X \\ \sqrt{\lambda_2} L_p^{\mathrm{T}} \end{pmatrix}, y^*_{(n+p)} = \begin{pmatrix} y \\ 0 \end{pmatrix}. \tag{13}$$

The LqCP estimator is defined as

$$\hat{\beta} = \arg\min_{\beta} \left\| y^* - X^* \beta \right\|_2^2 + \lambda_1 \left\| \beta^* \right\|_q^q. \tag{14}$$

Let

$$\begin{aligned} J(\beta) &= \left\| y - X\beta \right\|_2^2 + \lambda_1 \left\| \beta \right\|_q^q + \lambda_2 \left\| \beta \right\|_2^2 \\ &= \left[y^{\mathrm{T}} y + \beta^{\mathrm{T}} X^{\mathrm{T}} X \beta - 2 y^{\mathrm{T}} X \beta \right] + \lambda_2 \beta^{\mathrm{T}} W \beta + \lambda_1 \left\| \beta \right\|_q^q \\ &= \left[y^{\mathrm{T}} y + \beta^{\mathrm{T}} \left(X^{\mathrm{T}} X + \lambda_2 L L^{\mathrm{T}} \right) \beta - 2 y^{\mathrm{T}} X \beta \right] + \lambda_1 \left\| \beta \right\|_q^q \\ &= \left\| y^* - X^* \beta \right\|_2^2 + \lambda_1 \left\| \beta \right\|_q^q. \end{aligned} \tag{15}$$

Note that the sample size in the augmented problem is $n+p$ and X^* has rank p. As described in the paper of Marjanovic and Solo [15] in 2014, L_q $(0<q<1)$ sparsity penalizing linear regression can make a better sparsity result and a quick rate of convergence than L_1 norm. Whats more, they proposed a new algorithm called Cyclic Descent and have proved that minimizing the problem of sparsity penalized L_q in linear regression has global minimizers. In the base of that, the algorithm is used to solve the Equation (15) and the procedure is as follows:

$$\begin{aligned} J(\beta) &= \left\| y^* - X^* \beta \right\|_2^2 + \lambda_1 \left\| \beta \right\|_q^q \\ &= \left[\sum_{i=1}^{n} \left(y_i^* - \sum_{j \neq k} x_{ij}^* \beta_j - x_{ik}^* \beta_k \right)^2 \right] + \lambda_1 \sum_{j \neq k} \left| \beta_j \right|^q + \lambda_1 \left| \beta_k \right|^q \\ &= [\sum_{i=1}^{n} \left(y_i^* - \sum_{j \neq k} x_{ij}^* \beta_j \right)^2 + \sum_{i=1}^{n} x_{ik}^{*2} \beta_k^2 - 2 \sum_{i=1}^{n} \left(y_i^* - \sum_{j \neq k} x_{ij}^* \beta_j \right) x_{ik}^* \beta_k] + \lambda_1 \sum_{j \neq k} \left| \beta_j \right|^q + \lambda_1 \left| \beta_k \right|^q \\ &= \sum_{i=1}^{n} x_{ik}^{*2} \left[\beta_k^2 + \frac{\sum_{i=1}^{n} \left(y_i^* - \sum_{j \neq k} x_{ij}^* \beta_j \right)^2}{\sum_{i=1}^{n} x_{ik}^{*2}} - 2 \frac{\sum_{i=1}^{n} \left(y_i^* - \sum_{j \neq k} x_{ij}^* \beta_j \right) x_{ik}^* \beta_k}{\sum_{i=1}^{n} x_{ik}^{*2}} \right] + \lambda_1 \sum_{j \neq k} \left| \beta_j \right|^q + \lambda_1 \left| \beta_k \right|^q \\ &= \sum_{i=1}^{n} x_{ik}^{*2} \left[\beta_k^2 - 2 z_k \beta_k + z_k^2 \right] + \lambda_1 \left| \beta_k \right|^q + C \end{aligned} \tag{16}$$

Here, the related values are defined that $z_k = \frac{\sum_{i=1}^{n} \left(y_i^* - \sum_{j \neq k} x_{ij}^* \beta_j \right) x_{ik}^*}{\sum_{i=1}^{n} x_{ik}^{*2}}$,

$C = \sum_{i=1}^{n} \left(y_i^* - \sum_{j \neq k} x_{ij}^* \beta_j \right) - \frac{\left[\sum_{i=1}^{n} \left(y_i^* - \sum_{j \neq k} x_{ij}^* \beta_j \right) x_{ik}^* \right]^2}{\sum_{i=1}^{n} x_{ik}^{*2}} + \lambda_1 \sum_{j \neq k} \left| \beta_j \right|^q$,

$\lambda^* = \frac{\lambda_1}{\sum_{i=1}^{n} x_{ik}^{*2}}$. Then,

$$\min_{\beta} J(\beta_k) = \sum_{i=1}^{n} x_{ik}^{*2} \left[\left(z_k - \beta_k \right)^2 + \lambda^* \left| \beta_k \right|^q \right]. \tag{17}$$

Through the whole procedure, calculating the model (12) has been transformed to (15). Then, the algorithm procedure is stated below:

Initialize with β^1 and calculate the initial residual $r^1 := y^* - X^*\beta^1$. Denote the iteration counter by m and the coefficient index of an iterate by k. Set $m = k = 1$, and:

(a) Calculate the adjusted gradient $z_k^m := z\left(\beta_{-k}^m\right)$ by:

$$\begin{aligned} z_k^m &= \frac{X_{(k)}^{*\mathrm{T}}\left(y^* - X^*\beta_{-k}^m\right)}{X_{(k)}^{*\mathrm{T}}X_{(k)}^*} \\ &= \frac{X_{(k)}^{*\mathrm{T}}\left(y^* - X^*\beta^m + X_{(k)}^{*\mathrm{T}}X_{(k)}^*\beta^n\right)}{X_{(k)}^{*\mathrm{T}}X_{(k)}^*} \\ &= \frac{X_{(k)}^{*\mathrm{T}}r^m}{X_{(k)}^{*\mathrm{T}}X_{(k)}^*} + \beta_k^m \end{aligned} \tag{18}$$

(b) Use (a) computing the map:

$$\tau\left(z_k^m, \beta_k^m\right) := \begin{cases} 0 & \text{if } \left|z_k^m\right| < h_{\lambda^*,q} \\ \operatorname{sign}\left(z_k^m\right) b_{\lambda^*,q} I\left(\beta_k^m \neq 0\right) & \text{if } \left|z_k^m\right| = h_{\lambda^*,q} \\ \operatorname{sign}\left(z_k^m\right)\bar{\beta} & \text{if } \left|z_k^m\right| > h_{\lambda^*,q} \end{cases} \tag{19}$$

where $b_{\lambda^*,q} = \left[2\lambda^*(1-q)\right]^{\frac{1}{2-q}}$, $h_{\lambda^*,q} = b_{\lambda^*,q} + \lambda^* q b_{\lambda^*,q}^{q-1}$ and $\bar{\beta} > 0$ satisfies $\bar{\beta} + \lambda^* q\bar{\beta}^{q-1} = \left|z_k^m\right|$. There are two solutions to this equation and $\bar{\beta} \in \left(b_{\lambda^*,q}, \left|z_k^m\right|\right)$ is the larger one. It can be computed from the iteration, $b_{\lambda^*,q} \geq \bar{\beta}^{(0)} \leq \left|z_k^m\right|$:

$$\bar{\beta}^{(t+1)} = \rho\left(\bar{\beta}\right) \text{ where } \rho\left(\bar{\beta}\right) = \left|z_k^m\right| - \lambda^* q\bar{\beta}^{q-1} \tag{20}$$

(c) Update β^m by:

$$\beta^{m+1} = \beta_{-k}^m + \tau\left(z_k^m, \beta_k^m\right)e_k \tag{21}$$

where e_k has a 1 in the k-th position and 0's in the rest.

(d) Update the residual r^m with:

$$\begin{aligned} r^{n+1} &= y^* - X^*\beta^{m+1} = y^* - X^*\left(\beta_{-k}^m + \beta_k^{m+1}\right) \\ &= \left(y^* - X^*\beta_{-k}^m - \beta_k^m x_{(k)}^*\right) + \beta_k^m x_{(k)}^* - \beta_k^{m+1}X^*e_k \\ &= r^m - \left(\beta_k^{m+1} - \beta_k^m\right)x_{(k)}^* \end{aligned} \tag{22}$$

(e) Update the iteration counter m by $m = m+1$.

(f) Update the coefficient index k by:

$$k = \begin{cases} p & \text{if } 0 \equiv m \bmod p \\ m \bmod p & \text{otherwise} \end{cases} \tag{23}$$

(g) Go to (a)

2.2.3. Selection of Tuning Parameters

In practice, it is important to select appropriate tuning parameters in order to obtain a good prediction precision or estimation precision. There are three parameters $(\lambda_1, \lambda_2, q)$ which need to be chosen. As mentioned in the previous section, how to choose a proper q is important, which depends on the nature

of data. If sparsity of model is the point of focus, smaller q tends to be proper. The aim of this paper is not only to research the sparsity of variables but also to study grouping effect about strongly correlated variables. Therefore, the best q $(0<q<1)$ should be chosen by experiment data and cross-validation.

Firstly, q $(0<q<1)$ is given a grid of values to be compared. The choice of (λ_1,λ_2) will be different for different q. For each fixed q, cross-validation is used to choose the best (λ_1,λ_2). Typically, given a grid of values for λ_2, 5-fold cross-validation is applied to getting the best λ_1 in terms of minimized mean squared error of models or prediction error for each fixed λ_2. In simulation studies, the mean square error are defined by

$$\text{MSE}=\left(\hat{\beta}-\beta\right)^{\mathrm{T}}E\left(X^{\mathrm{T}}X\right)\left(\hat{\beta}-\beta\right), \tag{24}$$

where $E\left(X^{\mathrm{T}}X\right)$ is approximated by the sample covariance matrix of X of out-of-sample data. The formula of MSE is mentioned by Tibshirani [3] can measure the estimation error of model for simulation studies. In real data, because of unknown true parameters, test error defined by

$$\text{Test Error}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{25}$$

can measure prediction error of model. At last, the chosen value (q,λ_1,λ_2) is the best to compare with other models.

3. Simulation Studies

In this section, simulation studies are presented the finite sample performance of LqCP. The results analysis are considered from variable selection ability, the estimation errors, grouping effect. Data is generated from the true model:

$$y=X\beta+\sigma\epsilon,\epsilon\sim N(0,1). \tag{26}$$

y is the response variable and X is an $n\times p$ matrix with p predictor vectors and n observations. ϵ is a random error vector with $E(\epsilon)=0$. β is p dimension parameters and σ expresses the volatility of y. Three methods in the simulations study: the elastic net (ENET), the adaptive elastic net (AENET) and the L1CP are listed to be compared. Because these methods have the ability of grouped variable selection. Data is divided into two data sets: training data and testing data. Training data is used to do model fitting and cross-validation. Testing data is used to evaluate the error of models. For each estimator $\hat{\beta}$, its estimation accuracy is measured by the mean squared error (MSE) in the testing data. The variables selection performance is gauged by C and IC, where C is the number of zero coefficients that are correctly estimated by zero and IC is the number of non-zero coefficients that are incorrectly estimated by zero. In addition, the algorithm's stopping criterion is $\left\|\beta^{n+1}-\beta^{n}\right\|_2\le 10^{-4}$. In these simulation studies, q is given five different values which are $q=\left(0.1,\frac{1}{3},\frac{1}{2},\frac{2}{3},0.9\right)$. The results in the following tables are chosen by comparing MSE among different q. ENET and AENET are computed by

"glmnet" package and L1CP is computed by "lars" package in R language .

Example 1. 100 data sets are generated with sample size 100 observations from the linear regression model $y = X\beta + \sigma\epsilon, \epsilon \sim N(0,1)$, The true regression coefficients vector β is $\beta = (3, -2, 1.5, 0, \cdots, 0)$. Different dimension levels are considered as below. Smaller value $\sigma = 2$ is more proper and can make all results stable through the simulation experiment. The correlation matrix is given by $\rho(x_i, x_j) = 0.5^{|i-j|}$. The observation sample size of training data is 70 and testing data is 30. Furthermore, considering the problem of sparsity, noise variables increase by increasing the number of zero in β to generate four different dimensional data. The dimensions of data are respectively 10 (30%), 30 (10%), 60 (5%), 100 (3%). The number of noise variables are respectively 7, 27, 57, 97. That is to say, this part concerns that how the results change with the increase of noise variables. The simulation results are in Table 1 and Table 2.

Example 2. This example is about the case of $p > n$. Similar to the example of Wu, Shen and Geyer (2009), corresponding to high dimensional simulation scenarios with correlated groups, large p and small n, the X_i are simulated from $N(0, \Sigma)$, where the *jk*-th element of Σ is $0.5^{|j-k|}$ and

$$\beta = (3,3,3,3,3,-2,-2,-2,-2,-2,1.5,1.5,1.5,1.5,1.5,1,1,1,1,1,0_{180}).$$

So there are 20 grouped relevant predictors and 180 noise predictors and $n = 100$. Also, 100 data sets are generated and split data 70/30 into two parts for training data and testing data.

Example 3. About the grouping effect, 100 sample size described by 40 predictors is considered. The true parameters are

Table 1. Median mean squared errors of four methods based on 100 data replications with the standard errors estimated by using the bootstrap with B = 500 resamplings on the 100 mean squared errors for different numbers of noise variables.

Methods	30% level	10% level	5% level	3% level
ENET	0.521 (0.0314)	0.957 (0.0722)	1.501 (0.1342)	1.998 (0.1744)
AENET	0.166 (0.0280)	0.242 (0.0288)	0.535 (0.0680)	0.618 (0.1467)
L1CP	0.495 (0.0462)	0.824 (0.0543)	1.007 (0.0463)	1.336 (0.0724)
LqCP	0.200 (0.0309)	0.248 (0.0323)	0.372 (0.0573)	0.382 (0.1472)

Table 2. Median number of C and median number of IC based on 100 data replications for different numbers of noise variables. The number of true noise variables are 7, 27, 57, 97.

	30% level		10% level		5% level		3% level	
Methods	C	IC	C	IC	C	IC	C	IC
ENET	2	0	18	0	44	0	79.5	0
AENET	7	0	26	0	57	0	97	0
L1CP	3	0	17	0	46	0	82	0
LqCP	7	0	27	0	57	0	97	0

$$\beta = (2,2,2,2,2,3,3,3,3,3,4,4,4,4,4,0_{25})$$

and $\sigma = 6$ which is selected to see the performance of models in bigger volatility. The predictors are generated as:

$$x_i = Z_1 + 0.2\epsilon_i, Z_1 \sim N(0,1) \text{ and } i = 1,\cdots,5$$

$$x_i = Z_2 + 0.2\epsilon_i, Z_2 \sim N(0,1) \text{ and } i = 6,\cdots,10$$

$$x_i = Z_3 + 0.2\epsilon_i, Z_3 \sim N(0,1) \text{ and } i = 11,\cdots,15$$

x_i are independently identically distributed $N(0,1)$ for $i = 16,\cdots,40$, where ϵ_i are independently distributed $N(0,1)$ for $i = 1,\cdots,15$. In this data, three equally important groups have pairwise correlation $\rho \approx 0.96$, and there are 25 pure noise features. Also, the data is split as 70 observations for training data and 30 observations for testing data.

From Table 1, It is anticipated that MSE of all methods increase with the increase of noise variables. In 30% and 10% levels, AENET is better than other methods. But in in 5% and 3% levels, LqCP has the minimized MSE, which indicates LqCP performs better in low sparsity levels. In addition, LqCP is much better than ENET and L1CP for all levels. Table 2 illustrates that LqCP can estimate all true zero coefficients and the number of incorrect selection for true non-zero coefficients is always 0 with the increase of noise variables. Especially relative to ENET and L1CP, both of them can not select all true zero coefficients in four circumstances.

In the case of $p > n$ showed in Table 3, LqCP performs best and are more stable. As expressed in the example, the number of true noise variables is 180 and the result from LqCP is 179, which achieves a better selection of zero coefficients than other methods. From Table 4, LqCP performs with the minimized MSE and the minimized standard error which also has a better variable selection ability for true non-zero coefficients and zero coefficients in case of high correlations $\rho \approx 0.96$. It states better estimation effect of parameters of LqCP for grouping effect than other methods.

4. Real Data Sets Experiment

This part is about the performances of LqCP for two real world data sets: the US Crime and Gasoline described by $p = 15$ and $p = 401$ explanatory predictors respectively. The dimension p of US Crime data set is smaller than the sample

Table 3. Median mean squared errors of four methods based on 100 data replications with the standard errors estimated by using the bootstrap with B = 500 resamplings on the 100 mean squared errors. The dimension is 200 and the number of true noise variables is 180.

Methods	Median of MSE	C	IC
ENET	3.080 (0.2201)	151	0
AENET	1.795 (0.1956)	177	0
L1CP	2.172 (0.1124)	153	0
LqCP	1.456 (0.0767)	179	0

Table 4. Median mean squared errors of four methods based on 100 data replications with the standard errors estimated by using the bootstrap with B = 500 resamplings on the 100 mean squared errors. The number of true noise variables is 25.

Methods	Median of MSE	C	IC
ENET	8.351 (0.5367)	17	1
AENET	7.870 (0.4619)	19	2
L1CP	9.937 (0.5615)	20	4
LqCP	6.535 (0.4605)	23	0

size ($n = 47$), while the number of variables of Gaoline exceeds largely the sample size $n = 69$. Because the true parameters in application is unknown and the concern is prediction accuracy of response variable. Test Error is mentioned in the above as the criterion comparing among models. The selection of q is also based on the minimized test error for $q = \left(0.1, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 0.9\right)$.

4.1. US Crime Data

This data set is taken from R package "MASS" which contains 47 observations and 15 variables as well as one response variable. Criminologists are interested in the effect of punishment regimes on crime rates which has been studied using aggregate data on 47 states of the USA for 1960. This data set contained the following 16 columns: percentage of males aged 14 - 24 (M), indicator variable for a Southern state (So), mean years of schooling (Ed), police expenditure in 1960 (Pol), police expenditure in 1959 (Po2), labor force participation rate (LF), numbers of males per 1000 females (M.F), state population (Pop), number of non-whites per 1000 people (NW), unemployment rate of urban males 14 - 24 (U1), unemployment rate of urban males 35 - 39 (U2), gross domestic product per head (GDP), income inequality (Ineq), probability of imprisonment (Prob), average time served in state prisons (Time) and the outcome is the rate of crimes in a particular category per head of population (y). The data is split 100 times with a training set of 24 observations and a test set of 23 observations. The results are listed in **Table 5**.

Clearly, **Table 5** shows that LqCP selects 7 variables and has the minimized test error, which performs better than other methods followed by AENET on Crime data. Although the standard error of Test Error is not the lowest, it is very close to the standard error of AENET. The selected variables are percentage of males aged 14 - 24 (M), mean years of schooling (Ed), police expenditure in 1960 (Pol), numbers of males per 1000 females (M.F), number of non-whites per 1000 people (NW), income inequality (Ineq), probability of imprisonment (Prob). These variables can also be selected by other methods. Seeing from prediction accuracy and sparsity effect of models, LqCP is the best.

4.2. Gasoline Data

This data set "Gasoline" comes from R package "pls". It is about infrared

spectrum, which contains 69 observations. Recently, infrared spectrum is based on the function of diffuse reflecting degree measured by interval 2 nm from 900 nm to 1700 nm. Gasoline data have 401 prediction variables and the correlations of variables are very high that are almost 0.99. Similarly, the data set is split 100 times into a training set of 40 observations and a test set of 29. The prediction results are reported in Table 6.

In the circumstance of high dimension ($p = 401$) and small sample observations size($n = 69$), LqCP is the winner in term of test error, which also gets the least number of variables. Significantly, in this application, the correlation of variables is very high and approaches 1. The result shows that AENET, L1CP and LqCP have similar variable selection effect but ENET is the worst. Therefore, this application proved the efficiency of LqCP from the aspect of $p > n$ and highly correlated variables.

5. Conclusion

In this paper, motivated by variable selection and grouped selection property in linear regression problems, a new method called LqCP is proposed, which is a regularization procedure based on the penalized least squares with a mix of L_q $(0 < q < 1)$ norm and correlation-based penalty. Firstly, this paper discusses the current models that have grouping effect including elastic net, adaptive elastic net and L1CP. Similar to them, LqCP can also encourage grouping effect, where strongly correlation among predictors tend to be in or out of the model together. Through the simulation studies in the above, LqCP has better performances in terms of variable selection ability with large numbers of noise

Table 5. US Crime Data-Median test errors of four methods based on 100 random splits with the standard errors estimated by using the bootstrap with B = 500 resamplings on the 100 test errors. The median number of selected variables by each method is also reported.

Methods	Median of Test Error	Median no. of selected variables
ENET	0.530 (0.0163)	13
AENET	0.534 (0.0143)	10
L1CP	0.537 (0.0180)	9
LqCP	0.518 (0.0145)	7

Table 6. Gasoline Data-Median test errors of four methods based on 100 random splits with the standard errors estimated by using the bootstrap with B = 500 resamplings on the 100 test errors. The median number of selected variables by each method is also reported.

Methods	Median of Test Error	Median no. of selected variables
ENET	0.0213 (0.0006)	32
AENET	0.0217 (0.0009)	10
L1CP	0.0225 (0.0010)	11
LqCP	0.0191 (0.0007)	8

variables, high dimension $p > n$ and grouped variable selection automatically for high correlations of variables. Additionally, two real data proved LqCP's efficiency from the aspects of $p < n$ and $p > n$ through comparing prediction error with other models. Moreover, the oracle property is important in statistics area. One of future works is to pay attention to prove the oracle property of LqCP and also LqCP can be expended to general regression like logistic regression, quantile regression to solve some regression problems or multi-calss classification problems especially for data with highly correlated variables.

References

[1] Fan, J. and Li, R. (2006) Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *Marta Sanz Solé*, 595-622.

[2] Breiman, L. (1996) Heuristics of Instability and Stabilization in Model Selection. *Annals of Statistics*, **24**, 2350-2383. https://doi.org/10.1214/aos/1032181158

[3] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B*, **58**, 267-288.

[4] Knight, K. and Fu, W.J. (2000) Asymptotics for Lasso-Type Estimators. *Annals of Statistics*, **28**, 1356-1378. https://doi.org/10.1214/aos/1015957397

[5] 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. https://doi.org/10.1198/016214501753382273

[6] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[7] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*: *Series B*, **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[8] Zou, H. and Zhang, H.H. (2009) On the Adaptive Elastic Net with a Diverging Number of Parameters. *Annals of Statistics*, **37**, 1733-1751. https://doi.org/10.1214/08-AOS625

[9] Ghosh, S. (2011) On the Grouped Selection and Model Complexity of the Adaptive Elastic Net. *Statistics and Computing*, **21**, 451-462. https://doi.org/10.1007/s11222-010-9181-4

[10] Bondell, H.D. and Reich, B.J. (2008) Simultaneous Regression Shrinkage, Variable Selection and Clustering of Predictors with OSCAR. *Biometrics*, **64**, 115-123. https://doi.org/10.1111/j.1541-0420.2007.00843.x

[11] EL Anari, M. and Mkhadri, A. (2014) Penalized Regression Combining the L_1 Norm and a Correlation Based Penalty. *Sankhya B*, **76**, 82-102. https://doi.org/10.1007/s13571-013-0065-4

[12] Tutz, G. and Ulbricht, J. (2009) Penalized Regression with Correlation Based Penalty. *Statistics and Computing*, **19**, 239-253. https://doi.org/10.1007/s11222-008-9088-5

[13] Frank, I.E. and Friedman, J.H. (1993) An Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-135. https://doi.org/10.1080/00401706.1993.10485033

[14] Fu. W.J. (1998) Penalized Regressions: The Bridge vs the Lasso. *Journal of Compu-*

tational and Graphical Statistics, **7**, 397-416.

[15] Marjanovic, G. and Solo, V. (2014) l_q Sparsity Penalized Linear Regression With Cyclic Descent. *IEEE Transactions on Signal Processing*, **62**, 1464-475.