2007-05-11

Predictor Selection in Linear Regression: L1 regularization of a subset of parameters and Comparison of L1 regularization and stepwise selection

Qing Hu
Worcester Polytechnic Institute

Abstract

Background: Feature selection, also known as variable selection, is a technique that selects a subset from a large collection of possible predictors to improve the prediction accuracy in regression model. First objective of this project is to investigate in what data structure LASSO outperforms forward stepwise method. The second objective is to develop a feature selection method, Feature Selection by L_1 Regularization of Subset of Parameters (LRSP), which selects the model by combining prior knowledge of inclusion of some covariates, if any, and the information collected from the data. Mathematically, LRSP minimizes the residual sum of squares subject to the sum of the absolute value of a subset of the coefficients being less than a constant. In this project, LRSP is compared with LASSO, Forward Selection, and Ordinary Least Squares to investigate their relative performance for different data structures. Results: simulation results indicate that for moderate number of small sized effects, forward selection outperforms LASSO in both prediction accuracy and the performance of variable selection when the variance of model error term is smaller, regardless of the correlations among the covariates; forward selection also works better in the performance of variable selection when the variance of error term is larger, but the correlations among the covariates are smaller. LRSP was shown to be an efficient method to deal with the problems when prior knowledge of inclusion of covariates is available, and it can also be applied to problems with nuisance parameters, such as linear discriminant analysis.

Acknowledgments

I am very grateful to my advisors, Prof. Ryung S. Kim, for spending the great deal of time on advising me. His help and guidance are invaluable. I would also like to thank my co-advisor, Prof. Jayson D. Wilbur, for listening to me patiently and providing important comments on this project. I have benefited greatly from studying Applied Statistics at Worcester Polytechnic Institute. Prof. Joseph D. Petruccelli and Prof. Balgobin Nandram's teaching and their help on my job hunting are highly appreciated. Finally, I would like to thank my family and my friends for their support and encouragement.

Contents

List of Figures

List of Tables

1. Introduction

Feature selection, also known as variable selection, is the technique that selects a subset of relevant predictors to improve the prediction accuracy in regression model. It is also widely used to reduce dimensions of data in discriminant analysis setting. For example, tens of thousands of predictors (e.g. genes) are available in microarray data sets. An initial reduction in the number of predictors must be performed before fitting regression models or discriminant models to predict the sample phenotypes.

When outcome is a continuous variable, popular approach is to model the mean of response by linear combinations of predictors:

$$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Here, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m)$ is predictor matrix with observed values of the m predictors, $\mathbf{y} = (y_1, y_2, \cdots y_n)'$ are the observed values of the response variable, $\boldsymbol{\beta}$ is the vector parameter of length , and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)$ is the error vector, where ε_i's are iid normal random variables with mean zero and constant variance σ^2. One of classical methods to estimate parameters in linear regression is ordinary least squares (OLS), in which parameter estimates are chosen to minimize the residual sum of squares (RSS). That is, $\hat{\beta}_0^o$ and $\hat{\boldsymbol{\beta}}^o$ are chosen as the solution to the following optimization problem:

$$\begin{pmatrix} \hat{\beta}_0^o \\ \hat{\boldsymbol{\beta}}^o \end{pmatrix} = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \left\| \mathbf{y} - (\beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta}) \right\|_2^2 = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij})^2 \right\}$$

With some linear algebra, the formula for estimate of parameters and mean of response are given by following formula:

$$\begin{pmatrix} \hat{\beta}_0^o \\ \hat{\boldsymbol{\beta}}^o \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad \hat{\mathbf{y}} = \hat{\beta}_0^o \mathbf{1} + \mathbf{X}\hat{\boldsymbol{\beta}}^o$$

Being able to reliably, and automatically, select variables in linear regression models has drawn the attention of both applied and theoretical statisticians for a long time. Traditional methods include best subset selection and stepwise selection. The latter includes the techniques of forward addition and backward elimination. Each technique chooses the model according to some kind of statistical criterion which measures how good the model fits the data.

Best subset selection finds a subset among all possible combination of predictor variables that gives the best fit to a given data. This method, however, is not practical for a large number of predictor variables. For example, if there are 50 predictor variables in the data set, then the total number of all possible subset is 2^{50}. Even a recently developed efficient algorithm, leaps and bounds (Furnival and Wilson, 1974), is suitable for the number of predictor variables as large as 40 (Hastie et al, 2001). Instead of selecting model from all possible subsets, forward stepwise selection sequentially adds the predictor to the model based on the criterion adopted, while backward elimination reverses the forward stepwise procedure: it sequentially drops the predictor from the full model until all insignificant predictors have been removed. Compared to the best subset selection, stepwise selection procedures are relatively cheap in terms of computation, but they do have drawbacks: because of the "one-at-a-time" nature of adding/dropping variables, it is possible to miss the "optimal" model among all possible subsets. Moreover, stepwise selection may seriously overstate significance of results (Faraway, 2005).

Tibshirani (1996) proposed a new method for variable selection *least absolute shrinkage and selection operator*, also called as LASSO. The LASSO imposes a L_1 regularization of the parameters on the OLS estimates. In other words, it minimizes the

residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant.

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{m} |\beta_j| \right\} \quad (1)$$

Ridge regression has a similar objective function. However, it imposes L_2 regularization of the parameters on the OLS estimates rather than L_1. It is L_1 regularization that makes it possible for LASSO to shrink some coefficients and set others to zero as well, and hence retain both prediction accuracy and the purpose of variable selection.

To find the lasso solutions, change (1) to its equivalent problem:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \beta_j x_{ij})^2 \right\} \text{ subject to } \sum_{j=1}^{m} |\beta_j| \le t . \quad (2)$$

There is a one-to-one correspondence between the parameters λ in (1) and t in (2).

First, let $t \ge 0$ to be fixed. The constraint in (2) in effect includes 2^m linear inequality constraints, which correspond to the 2^m different possible signs for the β_j s. Due to the large number of constraints, Tibshirani (1996) suggested to solve (2) by starting from the overall least squares estimate, introducing the inequality constraints sequentially, searching a feasible solution satisfying Kuhn-Tucker conditions. Efron et al (2004) proposed yet another model selection algorithm, Least Angle Regression (LARS), which can be modified to solve the whole LASSO path simultaneously. It will be introduced in detail in the next chapter.

The first objective of this project is to investigate in what data structure LASSO outperforms forward stepwise method. Four simulation examples given by Tibshirani (1996) shows that in most cases LASSO performs better than best subset selection. But our simulation results indicate that for moderate number of small sized effect, 1). forward

stepwise outperforms LASSO in both mean squared error (*ME*) and model selection index (*MI*), when the variance of model error term is small, regardless of the correlations among the covariates; 2). forward stepwise also works better than LASSO in *MI*, when the variance of the error term is reasonably large but the correlations among the covariates are relatively small. Both *ME* and *MI* will be defined in Section 4.

The second objective of this project is to extend LASSO model to a more general L_1 regularized optimization problem in linear regression. That is, to minimize the residual sum of squares subject to the sum of absolute value of a subset of parameters being less than a constant, which refers to feature selection by L_1 regularization of subset of parameters (LRSP). The motivation comes from the following two considerations. First, investigators often have prior knowledge of true predictors for response (i.e., some of predictors need to be always kept in the model). Take the diabetes data used by Efron et. al. (2004) for example, 442 diabetes patients were measured on 10 baseline covariates: age, sex, body mass index, average blood pressure, glucose, and other five blood serum measurements. The response is a measure of disease progression one year after the baseline. It is known that glucose is a feature that contributes to diabetes for sure. Second, when nuisance parameters exist, investigators often intend to focus covariate selection only on the parameters of interest. Recall the gene expression example mentioned at the beginning of this Section, the parameters of interest are the differences between the genes of tumor cells and those of normal cells, while the genes of normal cells are actually nuisance parameters.

The structure of this report is as follows. In Section 2 we mathematically describe our problem and introduce the relevant methodology in detail. The algorithm and simulation results are provided in Section 3 and 4, respectively. In Section 5 we describe how to apply our method to linear discrimination analysis (LDA) for two-group case where the means of one group are nuisance parameters. Conclusions and future

research topics are discussed in Section 6.

2. L1 regularization of a subset of parameters (LRSP)

In this Section we provide formulations and brief descriptions of the new method, LRSP. Two model selection criteria that will be used in this project are also discussed. First, we define notations used in this report.

Notation

Each $\mathbf{x}_j$ is a predictor vector and there are *n* observations. *m* is the number of predictors. $\mathbf{X}_1 = (\mathbf{x}_{s1}, \mathbf{x}_{s2}, \cdots, \mathbf{x}_{sp})$ is a subset predictor matrix where $\{\mathbf{x}_{s1}, \mathbf{x}_{s2}, \cdots, \mathbf{x}_{sp}\}$ is a subset of predictor vectors $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$. $\boldsymbol{\beta}_1$ is the vector parameter of length p, which corresponds to $\mathbf{X}_1$. $\mathbf{X}_2$ is the matrix which is constructed by the remaining (m-p) predictor vectors. $\boldsymbol{\beta}_2$ is the vector parameter of length $(m-p)$, which corresponds to $\mathbf{X}_2$. β_0 is the intercept. $\mathbf{1}$ is a vector of 1's of length *n*. $\mathbf{I}_n$ is the $n \times n$ identity matrix. $\hat{\boldsymbol{\beta}}^o$ and $\hat{\beta}_0^o$ are the OLS estimate of $\boldsymbol{\beta}$ and β_0, respectively. Residual sum of squares is defined as following.

$$RSS = \begin{cases} \left\| \mathbf{y} - (\beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta}) \right\|_2^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij})^2, \text{ for unstandardized data,} \\ \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \beta_j x_{ij})^2, \text{ for standardized data.} \end{cases}$$

L_1 Constraint on Subset of Parameters (LRSP)

We extend LASSO model by incorporating prior knowledge part of predictors that should be always in the model:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\left\{\left\|\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1-\mathbf{X}_2\boldsymbol{\beta}_2\right\|_2^2+\lambda\left\|\boldsymbol{\beta}_2\right\|_1\right\} \qquad (3)$$

That is, penalize only a subset of coefficients by L_1 regularization. Analytical solution to this problem is difficult, if not impossible. However, a simple modification of LARS algorithm can solve this problem. Details of the algorithm and simulation results are left for Section 3 and 4, respectively.

3. Algorithm

Following is a slight modification of LARS and LASSO algorithm (Efron 2004) that solves our problem (3).

LRSP algorithm

Step 1. Standardizes **X** and **y** by centering and scaling $\mathbf{x}_j$, j = 1, …, m; and centering **y**.

Step 2. Fit the OLS to data ($\mathbf{X}_1$, **y**), and obtain the estimate $\hat{\boldsymbol{\beta}}_1^0$. Take the residual $\mathbf{r}_0 = \mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1^0$. Calculate $\hat{\mathbf{c}}_1 = \mathbf{X}^{\mathbf{T}}\mathbf{r}_0$, the correlation between **X** and $\mathbf{r}_0$, $\hat{C}_1 = \max_j\left\{\left|\hat{c}_{1j}\right|\right\}$ and the active set $S_1 = \left\{j:\left|\hat{c}_{1j}\right| = \hat{C}_1\right\} \cup S_0 = \left\{\hat{j}\right\} \cup S_0$, where S_0 is the index set of $\mathbf{X}_1$ .Increase the coefficient in the direction of OLS**,** $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1^0 + \gamma_1 A_1 \mathbf{S_1}\left(\mathbf{X}_{1'}^T\mathbf{X}_{1'}\right)^{-1}\mathbf{X}_{1'}^T\mathbf{r}_0 / \hat{C}_1$, where $A_1 = (1^T\mathbf{S}_1\left(\mathbf{X}_{1'}^T\mathbf{X}_{1'}\right)^{-1}\mathbf{X}_{1'}^T\mathbf{r}_0 / \hat{C}_1)^{-\frac{1}{2}}$, $\mathbf{X}_{1'}$ is the matrix that includes $\mathbf{X}_1$ and $\mathbf{x}_{\hat{j}}$, and

$$\gamma_1 = \min_{j\neq\hat{j}}^{+}\left\{\frac{\hat{C}_1-\hat{c}_{1j}}{A_1-a_{1j}}, \frac{\hat{C}_1+\hat{c}_{1j}}{A_1+a_{1j}}\right\}. \qquad (4)$$

$\min^+$ indicates that the minimum is taken over only positive components within each choice of j. Take residuals $\mathbf{r}_1 = \mathbf{y} - \mathbf{X}_{1'}\hat{\boldsymbol{\beta}}_1$ along the way. The value of γ_1 ensures

that some new index $\tilde{j}$ joins the active set and $\mathbf{x}_{\tilde{j}}$ has as much correlation with $\mathbf{r}_1$ as $\mathbf{x}_{\hat{j}}$ has, where $\tilde{j}$ is the minimizing index in (4). The new active set for next step is $S_2 = S_1 \cup \{\tilde{j}\}$. In fact, LRSP estimate $\hat{\boldsymbol{\beta}}_1$ have shrunk form the OLS estimate $\hat{\beta}_1^O$.

Step 3. Suppose *k* predictors are in the model, $\hat{\boldsymbol{\mu}}_{\mathbf{k}} = \mathbf{X}_k \hat{\boldsymbol{\beta}}_k$ and the new active set is S_{k+1}. Define the matrix $\mathbf{X}_{S(k+1)} = (\cdots s_{(k+1)j}\mathbf{x}_j \cdots)_{j \in S_{k+1}}$, where $s_{(k+1)j} = \text{sign}(\hat{c}_{(k+1)j})$ for $j \in S_{k+1}$ and $\hat{\mathbf{c}}_{k+1} = \mathbf{X}^T(\mathbf{y} - \hat{\boldsymbol{\mu}}_k)$, the correlation between $\mathbf{X}$ and residual $\mathbf{r}_k$. Let $\mathbf{G}_{k+1} = \mathbf{X}_{S(k+1)}^T \mathbf{X}_{S(k+1)}$ and $A_{k+1} = (\mathbf{1}_{k+1}^T \mathbf{G}_{k+1}^{-1} \mathbf{1}_{k+1})^{-\frac{1}{2}}$, then the unit vector $\mathbf{u}_{k+1} = \mathbf{X}_{S(k+1)} \mathbf{w}_{(k+1)}$, where $\mathbf{w}_{k+1} = A_{k+1} \mathbf{G}_{k+1}^{-1} \mathbf{1}_{k+1}$, makes equal angles, less than 90°, with the active predictors: $\mathbf{X}_{S(k+1)}^T \mathbf{u}_{k+1} = A_{k+1} \mathbf{1}_{k+1}$. Let $\mathbf{a}_{k+1} = \mathbf{X}^T \mathbf{u}_{k+1}$ and $\hat{C}_{k+1} = \max_j \{|\hat{c}_{(k+1)j}|\}$, then the new estimate with *k*+1 covariates is $\hat{\boldsymbol{\mu}}_{\mathbf{k+1}} = \hat{\boldsymbol{\mu}}_k + \gamma_{k+1} \mathbf{u}_{k+1}$ where

$$\gamma_{k+1} = \min_{j \in S_{k+1}}{}^{+} \left\{ \frac{\hat{C}_{k+1} - \hat{c}_{(k+1)j}}{A_{k+1} - a_{(k+1)j}}, \frac{\hat{C}_{k+1} + \hat{c}_{(k+1)j}}{A_{k+1} + a_{(k+1)j}} \right\}. \qquad (6)$$

The minimizing index joins the next active set S_{k+2}. It is the choice of equiangular vector $\mathbf{u}_{k+1}$ (which actually is $\mathbf{X}_k$'s joint least squares direction) and step size γ_k that make LARS less greedy than forward selection and more efficient in computation. The regression vector $\hat{\boldsymbol{\beta}}_{k+1}$ is given by $\hat{\boldsymbol{\beta}}_{k+1} = \hat{\boldsymbol{\beta}}_k + \gamma_{k+1} \mathbf{S}_{k+1} \mathbf{w}_{k+1}$, where $\mathbf{S}_{k+1}$ is the diagonal matrix with diagonal elements $s_{(k+1)j}$, for $j \in S_{k+1}$. The components of $\hat{\boldsymbol{\beta}}_{k+1}$ are:

$$\hat{\beta}_{(k+1)j} = \hat{\beta}_{kj} + \gamma_{k+1} s_{(k+1)j} w_{(k+1)j}, \text{ for } j \in S_{k+1}. \qquad (7)$$

Write (7) in another way:

$$\hat{\beta}_{kj}(\gamma) = \hat{\beta}_{kj} + \gamma s_{(k+1)j} w_{(k+1)j}$$

Still, the same derivation as in LARS algorithm, the LRSP estimates, $\hat{\beta}_{kj}$, have the same sign as the correlation $\hat{c}_{(k+1)j}$, for $j \in S_k$:

$$\text{sign}(\hat{\beta}_{kj}) = \text{sign}(\hat{c}_{(k+1)j}) = s_{(k+1)j}, \text{ for } j \in S_k. \quad (8)$$

Since $\hat{\beta}_{(k+1)j}$ will change sign from $\hat{\beta}_{kj}$ at

$\gamma_j = -\hat{\beta}_{kj} / s_{(k+1)j} w_{(k+1)j}$

Let

$$\tilde{\gamma} = \begin{cases} \min\{\gamma_j\}, \gamma_j > 0 \\ +\infty, \text{if no } \gamma_j > 0. \end{cases} \quad (9)$$

And $\tilde{\gamma} = \gamma_{\tilde{j}}$. If $\tilde{\gamma} < \gamma_{k+1}$ and $\tilde{j}$ is in the index set of X_2, for any γ between $\tilde{\gamma}$ and γ_{k+1}, $\hat{c}_{(k+1)\tilde{j}}(\gamma)$ does not change sign within such a single LRSP step since it is a continuous function but $\hat{\beta}_{k\tilde{j}}(\gamma)$ does. Thus the sign restriction (8) must be violated. As a result, stop the ongoing LRSP step at $\gamma = \tilde{\gamma}$ and drop $\tilde{j}$ from the active set.

Step 4. Repeat Step 3 until all predictors are in the model. Similar to LARS algorithm, estimates $\hat{\boldsymbol{\mu}}_k$ always approaching but never reaching the OLS estimates $\hat{\boldsymbol{\mu}}_k^{\mathrm{O}}$. This last step *m*, however, is an exception. S_m contains all predictors, thus (6) is not defined. Thus choose $\gamma_m = \hat{C}_m / A_m$ such that $\hat{\boldsymbol{\mu}}_m = \hat{\boldsymbol{\mu}}_m^o$ and $\hat{\boldsymbol{\beta}}_m = \hat{\boldsymbol{\beta}}_m^o$, the OLS estimate for the full set of *m* predictors.

If we do not enforce any prior knowledge of variables, and start with the intercept, this algorithm generates regular LASSO solution. Moreover, if we do not shrink the estimates of the parameters at each step, and do not drop any covariates either, LRSP becomes forward selection algorithm. Figure 1 and 2 show the examples of LASSO and LRSP solution paths for diabetes data used by Efron et. al (2004).

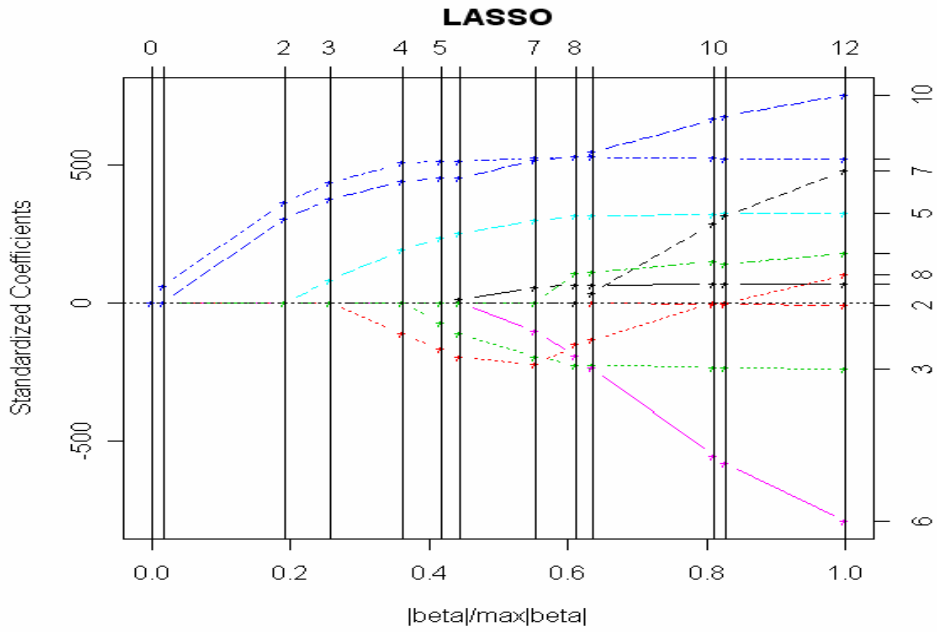

Figure 1. LASSO paths for diabetes data. The horizontal axis refers to the sum of absolute value of the estimates of coefficients divided by that of OLS estimates. The vertical axis refers to the estimates of coefficients for standardized data. The numbers on the top of the graph represent the LASSO steps.

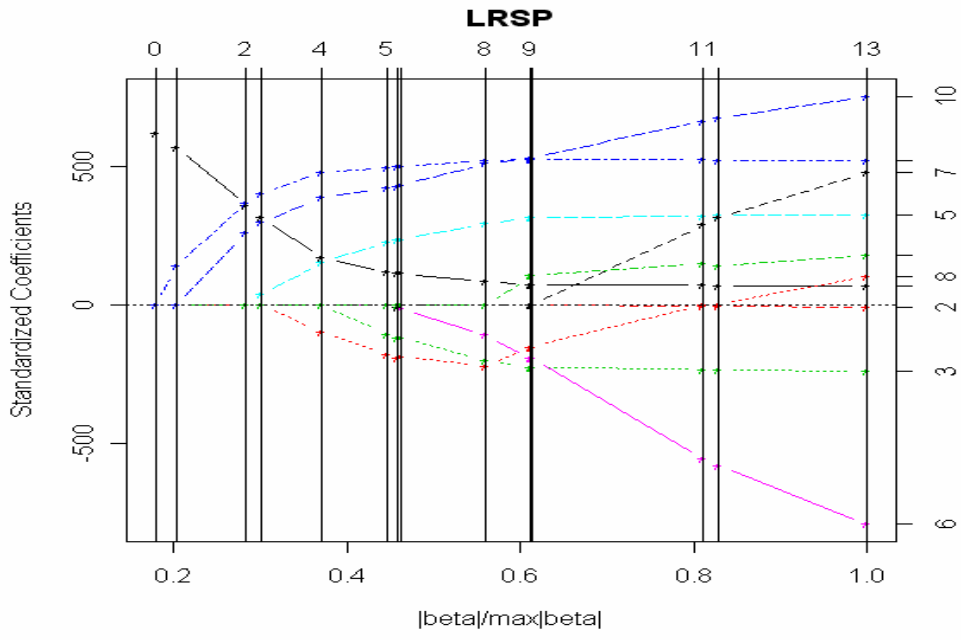

Figure 2. LRSP paths for diabetes data. Let glucose as the predictor of prior knowledge, and define constraint subset as other 9 covariates.

4. Simulations

In this section, first, we want to investigate when fix the setting of true parameters, how correlations (ρ) among predictors and the variance (σ^2) of error term of linear model will affect the performance of LASSO and forward selection. Second, once the big picture of their performance areas are obtained, we choose two representative pairs of (ρ, σ) to compare LASSO with LRSP, forward stepwise, and the full OLS estimates. All variable selection procedures discussed so far are implemented according to a specified criterion. We choose AIC_p and Cross-Validation as the criteria in this project. The subscription p indicates that there are p predictors in the model, p =1, …, m. For linear Gaussian models, $AIC_p = n \ln RSS_p - n \ln n + 2p$ is Akaike's information criterion. Models are searched by small values of AIC_p. Following is the description of K-fold Cross-Validation. Divide the data into K parts, equal (or close to) in size. For each part, use the rest of the data as the training set to fit the model, and that part as the test set to calculate the prediction error. Repeat for each part and average the result to obtain the CV error for this model. Choose the model with the smallest CV error among all models with different subsets of predictors.

4.1. Investigate in what data structure LASSO outperforms forward stepwise method

Here we compare LASSO with forward selection in both prediction accuracy and the performance of variable selection through two measurements, *ME* and *MI*. *AIC* is used

as a model selection criterion for both *ME* and *MI* comparison for different pairs of (ρ, σ).

300 data sets of 100 observations are generated, with 18 predictor variables.

$$\begin{pmatrix} x_{i1} & \cdots & x_{i,18} \end{pmatrix}^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}_X), i = 1,\ldots,100.$$

where

$$\boldsymbol{\Sigma}_X = \begin{pmatrix} 1 & \cdots & \cdots & \cdots & \rho & \cdots & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 & \cdots & \rho & 0 & 0 & 0 \\ \vdots & \cdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ \rho & \cdots & 0 & \cdots & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \rho & \cdots & 0 & \cdots & 1 & 0 & 0 & 0 \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 & 1 & 0 \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & 1 \end{pmatrix}_{18\times 18}$$

$\boldsymbol{\beta}$ = (1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0).

The response is obtained from the model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$. In the error term ε is a standard normal vector and the standard deviation σ is a parameter to be changed in the simulation. As Tibshirani (1996) did, we use mean squared error (*ME)* as a measure of prediction error. *ME* is estimated by $\hat{ME} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

To compare the performance of LASSO with forward stepwise in model selection, we define *MI* (model selection index) as:

$$MI = \frac{18 - a - b}{18}$$

a is the number of true covariates not selected and *b* is the number of false covariates selected. Thus if the model selected by LASSO or forward stepwise is exactly the true model, then *MI* = (18-0-0)/18 = 1; if the method select a completely wrong model, say, select the latter 9 covariates for the true $\boldsymbol{\beta}$ above, then *MI* = (18-9-9)/18 = 0. Otherwise, *MI* is between 0 and 1. Higher *MI* indicates that the corresponding method performs better in variable selection. We compute the ratios of *ME* LASSO to *ME* forward selection, and *MI* LASSO to *MI* forward selection, respectively. If the ME ratio is greater than 1, it means that forward selection works better than LASSO; if the MI ratio is greater than 1, it indicates that LASSO outperforms forward selection. Then we draw the heatmaps (Figure 3 and 4) of the ratios to illustrate the comparison of *ME* and *MI* of LASSO with that of forward selection. Figure 3 indicate that that forward selection performs better than LASSO in *ME* when the variance of model error term σ is smaller than 3. Figure 4 demonstrates that forward selection works much better than LASSO in variable selection when the variance of model error term σ is smaller than 4, regardless of the correlations among covariates, also performs better when σ is larger but ρ is relatively small.

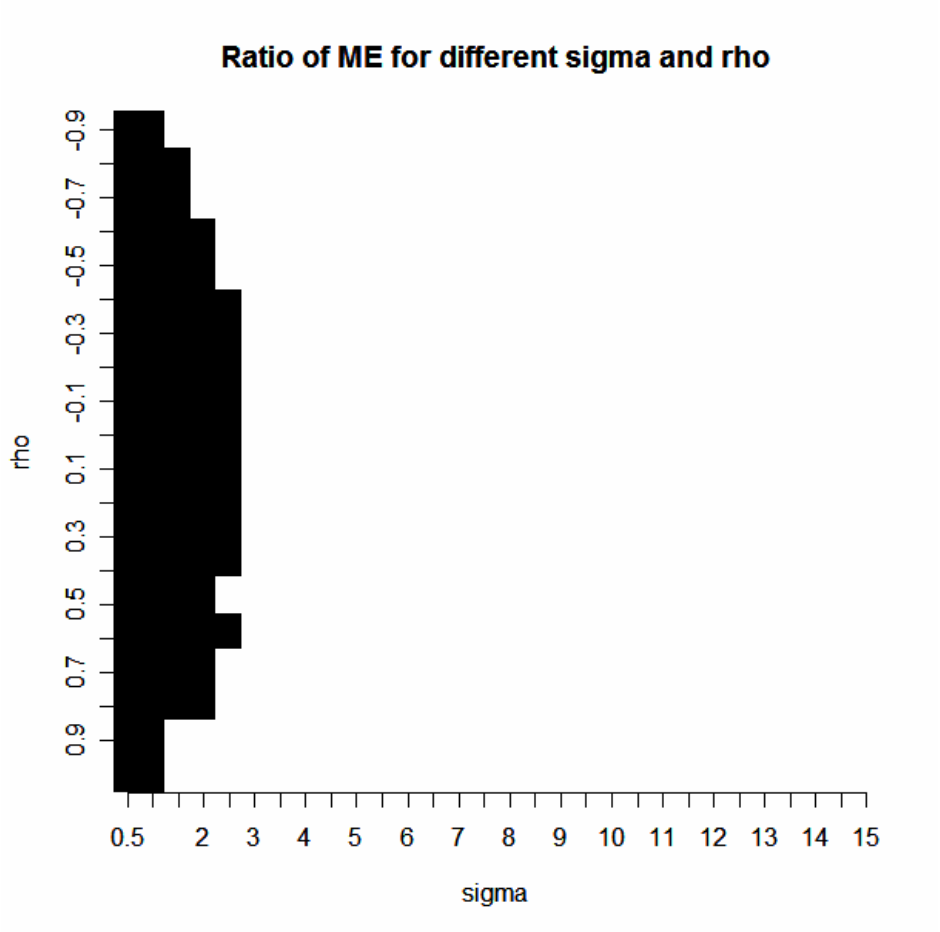

Figure 3. The heatmap of *ME* ratios. The vertical axis refers to ρ from -0.9 to 0.9, the horizontal axis refers to σ from 1 to 15. Black area indicates forward selection works better and white area LASSO works better.

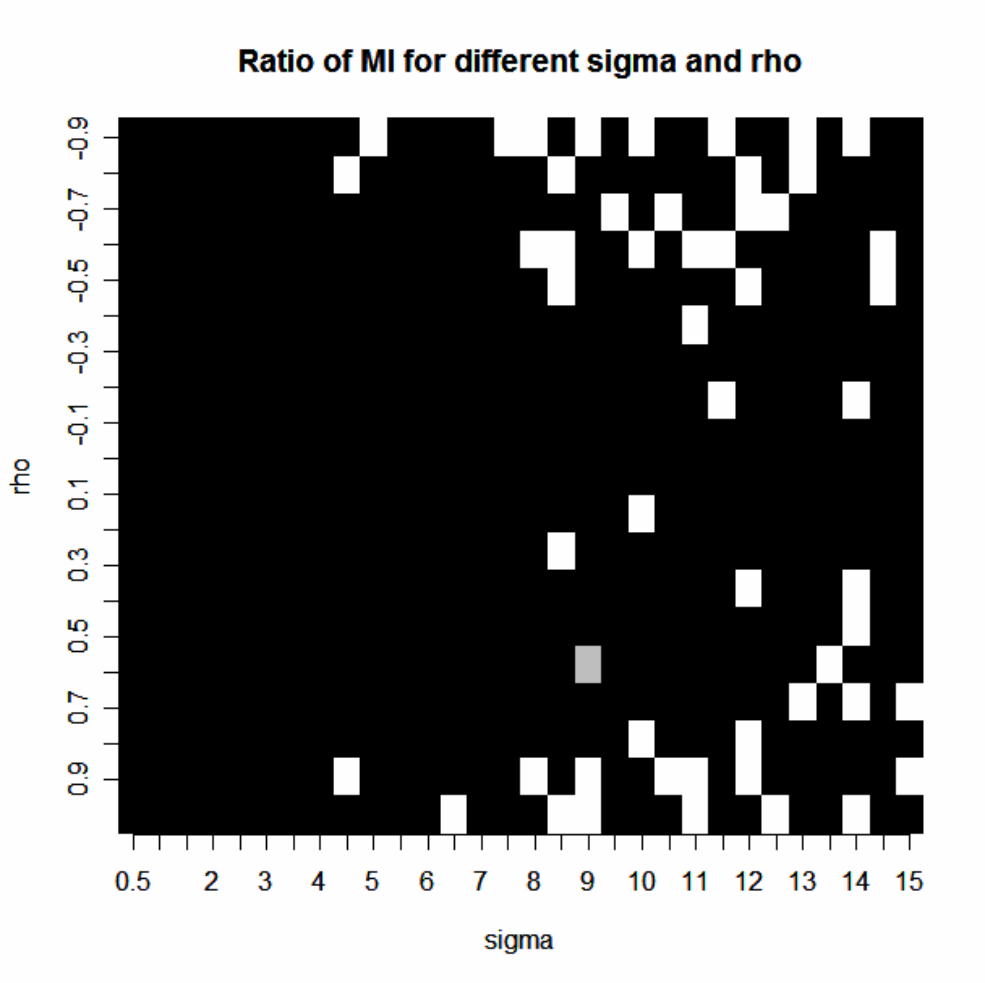

Figure 4. The heatmap of *MI* ratios. Black area indicates forward selection works better and white area LASSO works better.

4.2. Two representative cases

Now for the same data setting as in **4.1**, we choose two representative cases:

(ρ, σ) = (0, 1) and (ρ, σ) = (0.9, 15). Using 5-fold cross-validation (CV) as a criterion to select the model, and compare four methods, forward selection, LASSO, LRSP, and OLS, in their average of *ME*, *MI*, and CV, and the proportion of numbers of

covariates selected. Table 1 shows that in the case of low correlation among the predictors and low variance of the error term, forward stepwise performs the best in both MSE and variable selection, and LRSP works better than LASSO. The results in Table 2 show that when both correlation among the predictors and the variance of the error term of the linear model are very large, OLS performs best in MSE, followed by forward selection, LRSP, and LASSO. Regarding the variable selection, LRSP works best if prior knowledge of predictors is obtained, followed by LASSO, forward selection, and OLS.

Table 1. Comparison table when (ρ , σ) = (0, 1)

Method	ME	CV error (std)	Number of covariates										MI
			mean	Proportion									
				1	2	3	4	5	6	7	8	9	number of noises	
OLS	1.01	1.31 (0.21)	18	1	1	1	1	1	1	1	1	1	9	0.50
Forward	0.94	1.01 (0.17)	9.44	.98	.96	.98	.98	.98	1	.94	.94	.98	0.44	0.98
Lasso	1.04	1.28 (0.23)	13.86	.98	1	1	1	1	.98	1	1	.96	4.86	0.73
LRSP	1.03	1.27 (0.22)	13.06	1	1	1	1	1	.98	1	1	.96	4.06	0.77

The proportions of nine true covariates that have been chosen are listed, while the number of noises means the average number of false covariates that have been selected.

Table 2. Comparison table when (ρ , σ) = (0.9, 15)

Method	ME	CV error (std)	Number of covariates										MI
			mean	Proportion									
				1	2	3	4	5	6	7	8	9	number of noises	
OLS	50.15	308.05 (49.95)	18	1	1	1	1	1	1	1	1	1	9	0.50
Forward	13.31	214.51 (33.72)	0.6	.02	.02	.02	.06	.02	.04	.08	.06	.04	0.14	0.507
Lasso	10.72	230.57 (30.73)	2.84	.06	.06	.22	.16	.10	.12	.32	.22	.28	1.3	0.51
LRSP	14.73	238.48 (33.23)	5.78	1	1	1	.16	.28	.18	.28	.32	.26	1.26	0.68

The proportions of nine true covariates that have been chosen are listed, while the number of noises means the average number of false covariates that have been selected.

5. Application to LDA

In this Section, we will introduce the application of LRSP to Linear discriminant analysis. Even though we do not present data analysis here, LRSP can be easily implemented to apply L1 regularaization-based feature selection method to LDA. Such analysis is currently widely used in analysis of high dimensional data such as ones arising from microarray data. Considering two-group case, suppose that out of *p* (say, 10000) genes, we want to find significant genes that distinguish the observations $\mathbf{y}^{(1)}$ from cancer sample and $\mathbf{y}^{(2)}$ from normal sample. Assume the distribution for group *i* is multivariate normal with mean $\boldsymbol{\mu}^{(i)}$ and covariance $\Sigma^{(i)}$, $i = 1, 2$. Then we would like to test the difference between two group mean, $\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$. Furthermore we assume that the prior distributions are noninformative, and the covariance matrices are as following:

$$\boldsymbol{\Sigma}^{(i)} = \begin{pmatrix} \sigma_{i1}^2 & & \\ & \ddots & \\ & & \sigma_{ip}^2 \end{pmatrix}.$$

The model can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \Sigma\boldsymbol{\varepsilon}$ where $\mathbf{y}_1^{(1)}, \ldots, \mathbf{y}_n^{(1)} \overset{iid}{\sim} N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)})$, and $\mathbf{y}_1^{(2)}, \ldots, \mathbf{y}_m^{(2)} \overset{iid}{\sim} N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(2)})$.

$$\mathbf{y}_j^{(i)} = \begin{pmatrix} y_{j1}^{(i)} \\ \vdots \\ y_{jp}^{(i)} \end{pmatrix}, \boldsymbol{\mu}^{(i)} = \begin{pmatrix} \mu_{i1} \\ \vdots \\ \mu_{ip} \end{pmatrix}.$$

$$\mathbf{y} = \begin{pmatrix} y_{11}^{(1)} \\ \vdots \\ y_{n1}^{(1)} \\ \vdots \\ y_{1p}^{(1)} \\ \vdots \\ y_{np}^{(1)} \\ y_{11}^{(2)} \\ \vdots \\ y_{m1}^{(2)} \\ \vdots \\ y_{1p}^{(2)} \\ \vdots \\ y_{mp}^{(2)} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \mu_{11} \\ \vdots \\ \mu_{1p} \\ \mu_{21} - \mu_{11} \\ \vdots \\ \mu_{2p} - \mu_{1p} \end{pmatrix}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}), \Sigma = \begin{pmatrix} \sigma_{11}^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_{1p}^2 & & & \\ & & & \sigma_{21}^2 & & \\ & & & & \ddots & \\ & & & & & \sigma_{2p}^2 \end{pmatrix}.$$

Let $\Sigma^{-1} = \mathbf{W}$, then $\mathbf{Wy} = \mathbf{WX}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. This can be solved by iteratively reweighted least squares and LRSP with following substitutions:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \left\{ \left\| \tilde{\mathbf{y}} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2 \right\|_2^2 + \lambda \left\| \boldsymbol{\beta}_2 \right\|_1 \right\}$$

$$\boldsymbol{\beta}_1 = \begin{pmatrix} \mu_{11} \\ \vdots \\ \mu_{1p} \end{pmatrix}, \boldsymbol{\beta}_2 = \begin{pmatrix} \mu_{21} - \mu_{11} \\ \vdots \\ \mu_{2p} - \mu_{1p} \end{pmatrix}, \mathbf{WX} = (\mathbf{X}_1, \mathbf{X}_2), \tilde{\mathbf{y}} = \mathbf{Wy}.$$

6. Conclusions

This project is trying to investigate the performance of several model selection techniques with different data structure. The focus is on comparing LASSO with forward stepwise. Our results show that for moderate number of small sized effects, forward selection outperforms LASSO in both prediction accuracy (measured by *ME*) and the performance of variable selection (measured by *MI*) when the variance of model error term is smaller, regardless of the correlations among the covariates; forward selection also works better in the performance of variable selection when the variance of error term is larger, but the correlations among the covariates are smaller. In this project, we have developed a new approach (LRSP) that minimizes the residual sum of squares subject to L_1 constraint of subset of parameters. This approach can be applied to the problems when prior knowledge of some predictors is obtained. It can also be used in the problems with nuisance parameters where there is no interest in making inferences about those unknown nuisance parameters. The application to LDA is an example of this case. The present study may be extended to GLM as a future work. The algorithm of LRSP is developed by modifying LARS. Simulation results show that with prior knowledge LRSP performs better than LASSO in covariate selection.

References

Day, B.B., and Sandomire, M. Use of the discriminant function for more than two groups. *Journal of the American Statistical Association* 37 (220) (1942), pp. 461-472.

Efron, B.; T. Hastie; I. Johnstone; and R. Tibshirani. Least Angle Regression. *The Annals of Statistics* 32(2) (2004), pp. 407-499.

Faraway, J. *Linear Models with R*. Boca Raton: Chapman & Hall, 2005.

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(1936), pp. 179-184.

Furnival, G. M; and R.W. Wilson. Regressions by Leaps and Bounds. *Technometrics*, 16(4)(1974), pp. 499-511

Hastie, T.; R. Tibshirani; and J. Friedman. *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. New York: Springer Verlag, 2001.

Kutner, M. H.; C. J. Nachtsheim; and J. Neter. *Applied Linear Regression Models*. 4th ed. Boston: McGraw-Hill/Irwin, 2004.

Nigham, A.; and V. Aggarwal. The LPASSO Method for Regression Regularization. *Technical Report*. MIT, 2005.

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

Predictor Selection in Linear Regression:

L1 regularization of a subset of parameters and

Comparison of L1 regularization and stepwise selection

by

Qing Hu

A Project Report

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

May 2007

APPROVED:

_____

Dr. Ryung S. Kim, Major Advisor

_____

Dr. Jayson D. Wilbur, Co-advisor

_____

Dr. Homer Walker, Chair of Graduate Committee