Comparison of Some Variable Selection Techniques in Regression Analysis

December 12, 2020

Faruk Hossain
Department of Mathematical Sciences, UNLV

Abstract

Variable selection in linear models is essential for improved inference and interpretation, an activity which has become even more critical for high dimensional data. In this article, we provide a selective review of some classical methods including the forward selection method, the backward elimination method, the stepwise regression method, Akaike information criterion (AIC) Sakamoto et al. (1986), Bayesian information criterion (BIC) Raftery (1999), Mallow's Cp and risk inflation criterion, as well as regularization methods including Lasso, bridge regression, smoothly clipped absolute deviation, minimax concave penalty, adaptive Lasso, elastic-net, and group Lasso. We discuss how to select the penalty parameters. We also provide a review for some screening procedures for ultra high dimensions. We conduct simulation studies to make a comparison with these methods. We also analyze a real data set to further illustrate the comparison.

Keywords: Forward selection; Backward elimination; Stepwise regression; LASSO; Penalty method; variable selection.

1 Introduction

Variable selection is a fundamental problem in linear regression and has become increasingly important for many modern applications. During the past decade, a rich literature has been developed around this problem, especially for the case where large numbers of variables are collected and the number of variables exceeds the number of observations. There have been a growing concerns and controversies on the best variable selection techniques in regression analysis. There existed mixed results among them. Some were of the view that the use of forward selection method is better than the use of backward elimination method based of their findings, some were of the view that the method is outdated thereby bringing about better method of model building and variable selection. While some were of the view that their performance is the same. Fan et al. (2015) In a study stated that a good understanding of stepwise method require an analysis of the stochastic errors in the various stages of the selection problem, which is not a trivial task. Gunter et al. (2011) In a study to investigate four medical variable selection techniques: forward selection method, backward elimination method, stepwise regression method and the All subset combination method. He conducted an analysis on obtaining the best subset model among eight independent variables; chest, stay, Nratio, Culture, Facil, Nurse, Beds, and Census and his findings was that all the techniques gave the same subset models as best model except forward selection. Guyon and Elisseeff (2003) Argued that forward selection is computationally more efficient than backward elimination to generate nested subset of variables and cases when we need to get down to a single variable that work best on its own backward elimination would gotten rid of. Kira and Rendell (1992) Argued that weaker subset are found by forward selection, because the importance of variable is not assessed. This was illustrated with the use of an example where one variable separates the two classes better by itself than either of the two other ones taken alone and will therefore be selected first by forward selection. At the next step, when it is complemented by either of the two other variables the resulting class separation in two dimensions will not be as good as the one obtained jointly by the two variables that were discarded at the first step. And concluded that backward elimination method may outsmart forward selection by eliminating at first stage, the variable that by itself provides the best separation to retain the two variables that together perform best. Dale (2009) In a study to re-examine the scope of the literature addressing the weakness of variable selection methods and to re-enliven a possible solution of defining a better performing regression model. And after his study concluded that, finding the best possible subset of variables to put in a model has been a frustrating exercise. He said many variable selection methods exist and many statisticians know them, but few know they produce poorly performing models. He also said that resulting variable selection methods are a miscarriage of statistics because they are developed by debasing sound statistical theory to a misguided pseudo-theoretical foundation. And he quoted "I have reviewed the five widely used variable selection methods, itemized some of their weaknesses, and described why they are used. I have then sought to present a better solution to variable selection in regression: The Natural Seven-step Cycle of Statistical Modelling and Analysis. I feel that newcomers to Tukey's EDA need the seven-step Cycle introduced within the narrative of Tukey's analytic philosophy. Accordingly, I have embedded the solution within the context of EDA philosophy". Selena In a study of reviewing methods for selecting empirically relevant predictors from a set of N potentially relevant ones for the purpose of forecasting a scalar time series, using simulations to compare selected methods from the perspective of relative risk in one period ahead forecasts. One category consists of various penalized least squares methods, including the famous Lasso method by Tibshirani (1996) based on the convex ℓ_1 penalty for regularization, as well as non-convex penalties such as SCAD by FAN et al. (2001) and MCP by Zhou et al. (2010). The Lasso approach has also been extended to more sophisticated forms such as the group Lasso and graphical Lasso; see Tibshirani (2011) for a review. The other category consists of various Bayesian variable selection methods, such as stochastic search variable selection (SSVS) Yi et al. (2003) and Bayesian Lasso Park and Casella (2008). The two categories of methods are related in that the penalty terms correspond to specific Bayesian prior distributions. The penalized least squares approaches, especially the Lasso and its extensions, usually enjoy a computational advantage since the objective functions are convex and can be easily minimized. Despite the wide applicability of the linear regression model powered by modern variable selection tools, a single regression model can be inadequate if the data come from a heterogeneous population that consists of a number of different sub-populations with different characteristics. In this situation, it is possible that a separate linear regression model is needed for each sub-population. Moreover, the regression models in different sub-populations may use different subsets of predictor variables (or regressors, covariates) to explain the response variable. If the memberships of the observations are unobserved, then we naturally have a finite mixture model of linear regressions, where each mixture component is a linear regression model with its own subset of predictor variables. This gives rise to a variable selection problem that is more complex than that of a single linear regression model.

When our data that arise from a heterogeneous population Finite mixture models provide a flexible tool for modeling. They are used in many fields, including biology, genetics, engineering, and marketing. The book by Peel and McLachlan (2000) contains a comprehensive review of finite mixture models. When a random variable with a finite mixture distribution depends on certain covariates, we obtain a finite mixture of regression (FMR) model. Jacobs, Jordan, Nowlan, and Hinton (1991) and Jiang and Tanner (1999) have discussed the use of FMR models in machine learning applications under the term mixture of experts models. The books by Wedel and Kamakura (2012) and Skrondal and Rabe-Hesketh (2004), among others, contain comprehensive reviews on the applications of FMR models in market segmentation and the social sciences. Often, in the initial stage of a study many covariates are of interest, and their contributions to the response variable vary from one component to another of the FMR model. To enhance predictability and to give a parsimonious model, it is common practice to include only the important covariates in the model. The problem of variable selection in FMR models has received much attention recently. All-subset selection methods, such as the Akaike information criterion (AIC; Akaike 1973), the Bayes information criterion (BIC; Schwarz 1978), and their modifications, have been studied in the context of FMR models; for instance Wang et al. (1996) used AIC and BIC in finite mixture of Poisson regression models. However, even for FMR models with moderate numbers of components and covariates, all-subset selection methods are computationally intensive. In this article we are using Forward selection, Backward elimination, Stepwise and Shrinkage variable selection method (lasso) along with some other procedures and compare them. We also investigate methods for selecting tuning parameters adaptively and develop an EM algorithm for numerical computations. We want to see which variable selection is consistent. The performance of these method is studied theoretically and by simulations. Our simulations indicate that the lasso method is as good as or better than BIC at selecting correct models, with much less computational effort. The article is organized as follows. In Section 2 presents the formal selection method and lasso along with their identifiability. Section 3 illustrates the penalized likelihood-based approach for variable selection. Section 4 describes large sample properties with choosing tuning parameters. New method by simulation studies, where new method is compared with existing methods are shown in section 5. Section 6 discusses of a real data set to further illustrate this method. Finally Section 7 concludes with a brief discussion.

2 Regression Theory

Regression is a statistical tool for evaluating the relationship between one or more dependent variables $x_1, x_2, x_3, \dots x_n$ and a single continuous dependent variable Y. It is most often used when the independent variables are not controllable, that is, when collected in a sample survey or other observational studies. There are so many types of regression model, but just one of the regression models will be used for this study and that is the linear regression model.

2.1 Linear Regression Model:

A regression model is linear if and only if a variable is a function of another variable whose power equals ones. There are two types of linear regression, and they are; the simple linear regression and the multiple linear regression. For the purpose of the analysis, just the multiple regression will be considered.

2.1.1 Multiple Linear Regressions:

Multiple linear regression is a statistical analysis that fits a model to predict a dependent variable from some independent variables. Multiple regressions involve more than one independent variable. The relationship between the dependent and independent variable is expressed as follows:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 +, ... + \beta_k x_k + \epsilon_i \tag{1}$$

Where,

Y_i is the i-th response or dependent variable

X_i is the i-th independent variable

ϵ_i is the error term of the i-th observation, which is normally, independently distributed with mean zero and variance

$\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_k$ are the observation parameter.

2.1.2 Estimation of Parameters of the Model:

Unbiased estimates of the parameters $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_k$ can be obtained by several methods. The most widely used is the method of least squares. This means that the sum of the square's deviation of the observed values of Y from their expected value is minimized. In other words, by the method of least squares, sample estimates of $b_0, b_1, b_2, \dots .b_k$ of $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_k$ respectively are selected in such a way that $Q = \sum \epsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} -, ... - \beta_k x_{ik})^2$ is minimized. As in simple regression, we obtain estimates, $b_0, b_1, b_2, \dots .b_k$ of the regression coefficients by solving the following set of normal equations,

$$\sum Y_i = nb_0 + b_1 \sum x_{i1} + b_2 \sum x_{i2} + ... + b_k \sum x_{ik}$$

$$\sum X_{i1} Y_i = b_0 \sum X_{i1} + b_i \sum {X_{i1}}^2 + b_2 \sum X_{i1} X_{i2} + ... + b_k \sum X_{i1} X_{ik}$$

$$\sum X_{i2} Y_i = b_0 \sum X_{i2} + b_i \sum X_{i1} X_{i2} + b_2 \sum {X_{i2}}^2 + ... + b_k \sum X_{i1} X_{ik}$$

$$\sum X_{ik} Y_i = b_0 \sum X_{ik} + b_i \sum X_{i1} X_{ik} + b_2 \sum X_{i1} X_{i2} + ... + b_k \sum X_{ik}^2$$

Although these normal equations are obtained mathematically by finding estimates $b_0, b_1, b_2, \dots .b_k$ that would minimize equation Q, a simple procedure to remember in obtaining them is as follows: The usual regression equation is written down with $b_0, b_1, b_2, \dots .b_k$ as coefficients. The first normal equation is then obtained by summing each term of this regression equation. The second normal equation is obtained by multiplying every term in the regression equation by Xi1 and summing the result. The third normal equation is obtained by multiplying every term in the regression equation by Xi2 and summing the result; and so on. It will be too cumbersome to obtain separate expressions for the estimates $b_0, b_1, b_2, \dots .b_k$. Instead these coefficients are obtained by calculating the required sums from the data for the various combinations of $X_1, X_2, ...X_k$ and substituting this sum into the normal equations, which are solved simultaneously.

2.1.3 Assumptions of The Regression Analysis:

The following are the assumptions of the regression analysis above:

a) X values are fixed in repeated sampling.

b) Zero mean value of disturbance U_I. Given the value of X, the mean, or expected value of the random disturbance term U_I is zero. Symbolically, we have $E(\mu_i/X_i) = 0$.

c) Homoscedasticity or equal variance. Given the value of X, the variance of is the same for all observations. That is, the conditional variances of are identical.

d) No auto-correlation between the disturbances. Given any two X values, and $(i \neq j)$ is zero.

e) Zero co-variance between $_i$ and X_i , or $E(\mu_i X_i) = 0$

f) The number of observations n must be greater than the number of parameters to be estimated.

g) Variability in X values. The X values in each sample must not all be the same.

h) There is no perfect multicollinearity. That is, there are no perfect linear relationships among the explanatory variable.

i) The error term is normally distributed, that is, $\mu_i \sim N(0, \delta_u^2)$

2.1.4 Hypothesis testing:

In hypothesis testing there are two types: Null hypothesis and Alternative hypothesis. Null hypothesis is the hypothesis being tested. It is often formulated with the purpose of being rejected. It is stated as: $H_0 : \beta = 0$ which shows that the coefficients are the same. Alternative hypothesis is the hypothesis that contradict the null hypothesis. It is stated as $H_1 : \beta > 0$ or $H_1 : \beta < 0$ or $H_1 : \beta \neq 0$ Shows that the coefficients are not the same.

2.1.5 Homoscedasticity:

Observations are said to be homoscedastic if they have equal or constant variance. Since one of the assumptions of regression is that the residuals have constant variance, we would be using a scatter plot of the standardized predictors to reach a conclusion on homoscedasticity.

2.1.6 Autocorrelation:

The term autocorrelation may be defined as correlation between members of series of observations ordered in time, that is, time series data or space/ cross sectional data. The most celebrated test for detecting serial correlation is that developed by statisticians Durbin and Watson d statistic. A great advantage of the d statistic is that it is based on the estimated residuals, which are routinely computed in regression analysis. However, Durbin and Watson were successful in deriving a lower bound dl and a upper bound du such that if the computed d lies outside these critical values, a decision can be made regarding the presence of positive or negative serial correlation. Moreover, this limit depends only on the number of observations n and the number of explanatory variables. The limits, for n going from 6 to 200 and up to 20 explanatory variables have been tabulated by Dublin and Watson and the limits of 0 and 4.

a) Test of hypothesis: H_0 : There is no autocorrelation; H_1 : There is autocorrelation.

b) Decision: Reject the null hypothesis if the Durbin- Watson d statistics value falls outside the limit of d, that is, within the range of 0 and 4.

2.1.7 Level of Significance:

The level of significance is the difference between the percentage required and 100 percent. For instance, if 95 percent certainly is required, then the level of significance will be denoted as $\alpha = 0.05$. This is the probability of committing a type one error, while a type one error is simply rejecting a true null hypothesis.

2.1.8 Test for Model Adequacy:

Test of Hypothesis:
H_0: The model is not adequate;
H_1 : The model is adequate. Using a 5% level of significance.
a)Test Statistic:

$$F_{cal} = \frac{MS_{reg}}{MS_{res}}$$

b) Decision Rule: Reject H_0 if $F_{cal} > F$, accept if otherwise

2.1.9 Test for Parameter Significance:

This is simply a test of the significance of the individual parameters in the model.
a) Test of hypothesis: $H_0 : \beta = 0$ (The coefficient is not statistically significant); $H_1 : \beta \neq 0$ (The coefficient is statistically significant). Using a 5% level of significance $t_{cal} = \frac{\tilde{\beta_i}}{se(\tilde{\beta_i})} \sim t^{\alpha-2}_{n-k}$

b) Decision Rule: Reject H_0 if $t_{cal} > t_{tab}$ accept if otherwise.

2.1.10 Critical Region:

Critical region indicates the value of the test statistic that will imply rejection of the null hypothesis. It is also called the rejection region. Opposite of the critical region is the acceptance region, which indicates the value of the test statistics that will imply acceptance of the null hypothesis.

2.1.11 Multicollinearity:

Multicollinearity is used to denote the presence of of linear or near linear dependence among the explanatory variables. Multiple regression model with correlated explanatory variables indicate how well the entire bundle of predictors predicts the outcome of the variable, but it may not give valid results about any individual predictor or about which predictors are redundant with others. We have a perfect multicollinearity if the correlation between two independent variables are equal to +1 or -1, such that we have exact linear relationship if the following condition is satisfied: $a_1x_1 + a_2x_2 + a_3x_3 + ... + a_kx_k = 0$. But if the X variables are intercorrelated to the Y variables, we have: $a_1x_1 + a_2x_2 + a_3x_3 + ... + a_kx_k + v_i$. Where α is the constant, is the random term, and represents the independent variables with $i = 1, 2, 3, \ldots, k$.

Multicollinearity Diagnostics: Several techniques have been proposed for detecting multicollinearity, but here three techniques will be considered. The desirable characteristics of a diagnostic measure are that it directly reflects the degree of the multicollinearity problem and provide information helpful in determining which regressors are involved. We have; the examination of the correlation matrix, the variance inflation factors, the eigensystem analysis of X'X. Here the variance inflation factor was used to account for the effect of multicollinearity on the various subset regression model.

$$VIF = \frac{1}{1 - R^2_{ij}}$$

2.2 The LASSO estimator:

LASSO is a regularization and variable selection method for statistical models. We first introduce this method for linear regression case. The LASSO minimizes the sum of squared errors, with a upper bound on the sum of the absolute values of the model parameters. There are different mathematical form to introduce this topic, we will refer to the formulation used by Tibshirani (1997). The lasso estimate is defined by the solution to the l_1 optimization problem minimize$\left(\frac{||Y-X\beta||_2^2}{n}\right)$ subject to $\sum_{j=1}^{k} ||\beta||_1 < t$ where t is the upper bound for the sum of the coefficients. This optimization problem is equivalent to the parameter estimation that follows

$$\hat{\beta}(\lambda) = \underset{\beta}{argmin} \left(\frac{||Y - X\beta||_2^2}{n} + \lambda||\beta||_1 \right)$$

where $||Y - X\beta||_2^2 = \sum_{i=0}^{n}(Y_i - (X\beta)_i)^2, ||\beta||_1 = \sum_{j=1}^{k} |\beta_j|$ and $\lambda \geq 0$ is the parameter that controls the strength of the penalty, the larger the value of λ, the greater the amount of shrinkage. The relation between λ and the upper bound t is a reverse relationship. Indeed as t becomes infinity, the problem becomes an ordinary least squares and λ becomes 0. Vice-versa as t becomes 0, all coefficients shrink to 0 and λ goes to infinity. In this research paper we are going to use LASSO for its variable selection property. When we minimize the optimization problem some coefficients are shrank to zero, i.e. $\hat{\beta}(\lambda) = 0$, for some values of j (depending on the value of the parameter λ). In this way the features with coefficient equal to zero are excluded from the model. For this reason LASSO is a powerful method for feature selection while other methods (e.g. Ridge Regression) are not. As shown in Figure1 both Ridge Regression and LASSO methods find the first point where the least squares error function border touch the constraint region. For the LASSO method the constraint region is a diamond, therefore it has corners; this means that if the first point is in proximity of the corner, then it has one coefficient β_j equal to zero. While for the Ridge Regression method the constraint region is a disk, thus it has no corners and the coefficients can not be equal to zero.

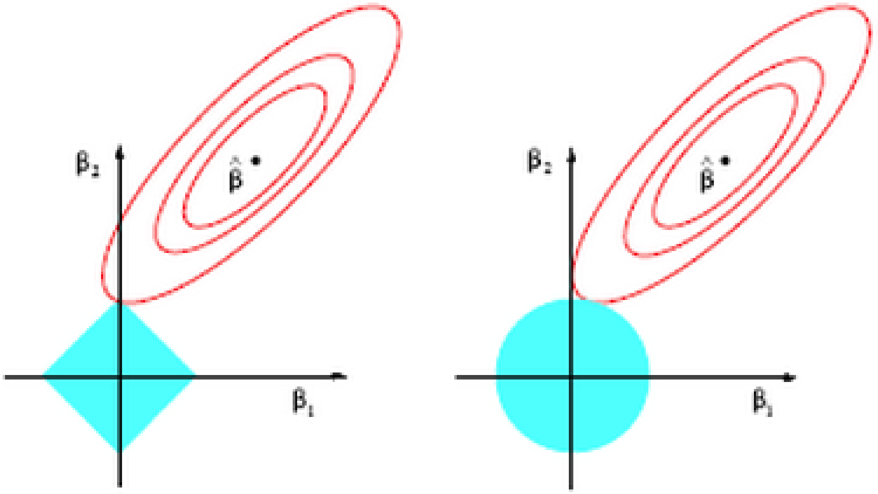

Figure 1: Graph for the LASSO Regression (left side) and Ridge (right side).The two areas are the constraint regions: the disk for the ridge regression,$\beta_1^2 + \beta_2^2 \leq t$, and the diamond for the LASSO, $|\beta_1| + |\beta_2| \leq t$ While the ellipses are the borders of the least squares error functions.

3 Variable selection Techniques:

It is desirable to consider regression models that employ a subset of the candidate regressor variables. To find the subset of variables to use in the final equation, it is natural to consider fitting models with various combinations of the candidate regressors. There are several computational techniques for generating subset regression models, but our concentration would be based on four of these techniques and they are; All possible Regressions, Direct Search on t, Forward Selection method, backward elimination method, and the stepwise regression.

3.1 All Possible Regressions:

This procedure requires that the analyst fit all regression equations involving one candidate regressor, two candidate regressors and so on. These equations are evaluated according to some suitable criterion and the best regression model selected. If we assumed that the intercept term $\beta_0$is included in all equations, if there are k candidate regressors, there are 2^k total equations to be estimated. For the appropriate model, we consider the adjusted R^2 which result is insensitive to the number of variables in the model making it appropriate for decision making in this method where we have to choose the best model combination from the various model combinations, but also the size of the model, that is, the number of variables in the model should also be taken into consideration because the more the variables the more the information obtained from these factors, which actually has a strong influence on the predicted value of independent variable. Though one needs to be careful as to the statement made with respect to the size, because the more the variable the higher the variance of the prediction.

3.2 Direct search on t:

The test statistics for testing $H_0 : \beta_j = 0$ for the full model with p=K+1 regressors is: $t_{k,j} = \frac{\tilde{\beta}_j}{se(\tilde{\beta}_j)}$. Regressors that contribute significantly to the full model will have a large $t_{k,j}$ and will tend to be included in the best p-regressor subset, where best implies minimum residual sum of squares or C_p. Consequently ranking the regressors according to decreasing order of magnitude of the $t_{k,j}, j = 1, 2, \ldots, k$ and then introducing the regressors into the model one at a time in this order should lead to the best or one of the best subset models for each p.

3.3 Forward Selection Method:

This technique begins with the assumption that there are no regressors in the model other than the intercept. An effort is made to find an optimal subset by inserting regressors into the model one at a time. The first regressor selected for entry into the equation is the one that has the largest simple correlation with the response variable y, it is also the regressor that will produce the largest F- statistic for testing significance of regression. This regressor is entered if the F-statistics exceeds a preselected F value, say F_{IN} (or F-to-enter). The second regressor chosen for entry is the one that now has the largest correlation with y after adjusting for the effect of the first regressor entered. We refer to these correlations as partial correlations. They are simple correlation between the residuals from the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$ and the residuals from the regressions of the other candidate regressors on x_1 , say$\hat{x} = \hat{\alpha_{0j}} + \hat{\alpha_{1j}} x_1$, j=2,3,...k. Here, the regressor with the highest partial correlation which also implies the largest F-statistic and if its F value exceeds F_{IN} then the regressor enters the model. In general, at each step the regressor having the highest partial correlation with y given the other regressor already in the model is added to the model if its partial F-statistic exceeds the preselected entry level FIN. This procedure terminates either when the partial F statistics at a point does not exceed or when the last candidate regressor is added to the model.

3.4 Backward elimination method:

Forward selection begins with no regressors in the model and attempt to insert variables until a suitable model is obtained. Backward eliminations attempt to find a good model by working in the opposite direction. That is, we begin with a model containing all K candidate regressors. Then the partial F-statistic partial F-statistic is compared with a preselected value, F_{OUT} (F−to−remove), for example, and if the smallest partial statistics is less than F_{OUT} that regressor is removed from the model.

3.5 Stepwise regression method:

This can be said to be a modification of the forward selection in which at each step all regressors entered the first model previously are regressed via the partial F-statistics. A regressor added at an earlier step may now be redundant because of the relationship between it and regressors in the equation. If the partial F-statistics for a variable is less than F_{OUT}, that variable is dropped from the model. This method requires two cut off values, F_{IN} and F_{OUT}. Some analyst prefers to choose $F_{IN} = F_{OUT}$,although this is not necessary. Frequently we choose $F_{IN} > F_{OUT}$, making it relatively more difficult to add a regressor than to delete one.

3.6 LASSO

LASSO is a regularization and variable selection method for statistical models. We first introduce this method for linear regression case. The LASSO minimizes the sum of squared errors, with a upper bound on the sum of the absolute values of the model parameters. There are different mathematical form to introduce this topic, we will refer to the formulation used by Buhlmann and van de Geer [1]. The lasso estimate is defined by the solution to the l_1 optimization problem
$minize\frac{||Y-X\beta||_2^2}{n}$ subject to $\sum_{j=1}^{k} ||\beta||_1$

where t is the upper bound for the sum of the coefficients. This optimization problem is equivalent to the parameter estimation that follows

$$\hat{\beta}(\lambda) = argmin(\frac{||Y + X\beta||_2^2}{n} + \lambda||\beta||_1)$$

where $||Y - X\beta||_2^2 = \sum_{i=0}^{n}(Y_i - (X\beta)_i)^2, |\beta_1| = \sum_{j=1}^{k} |\beta_j|$ and $\lambda \geq 0$ is the parameter that controls the strength of the penalty, the larger the value of λ, the greater the amount of shrinkage. The relation between λ and the upper bound t is a reverse relationship. Indeed as t becomes infinity, the problem becomes an ordinary least squares and λ becomes 0. Viceversa as t becomes 0, all coefficients shrink to 0 and λ goes to infinity. In this research paper we are going to use LASSO for its variable selection property. When we minimize the optimization problem some coefficients are shrank to zero, $\hat{\beta}_j = 0$, for some values of j (depending on the value of the parameter λ). In this way the features with coefficient equal to zero are excluded from the model. For this reason LASSO is a powerful method for feature selection while other methods (e.g. Ridge Regression) are not.

3.7 Adaptive LASSO

The lasso cannot be an oracle procedure. However, the asymptotic setup is somewhat unfair, because it forces the coefficients to be equally penalized in the l_1 penalty. We can certainly assign different weights to different coefficients. Let us consider the weighted lasso, known as Adaptive LASSO introduced by Zou (2006)

$$\hat{\beta}(\lambda) = argmin(\frac{||Y + X\beta||_2^2}{n} + \lambda \sum_{j=1}^{p} w_i ||\beta||_1)$$

where w is a known weights vector. The weights are data-dependent and cleverly chosen, then the weighted lasso can have the oracle properties.

3.8 SCAD

A non-concave penalty function referred to as the smoothly clipped absolute deviation (SCAD) by Fan and Li FAN et al. (2001). The SCAD penalty is given by

$$\mathrm{P}_{\lambda}^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leq \lambda \\ -\frac{|\beta_j^2| - 2a\lambda\beta_j + \lambda^2}{2(a-1)} & \text{if } \lambda < |\beta_j| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } \beta_j > a\lambda \end{cases}$$

SCAD can produce sparse set of solution and approximately unbiased coefficients for large coefficients. The solution to the SCAD penalty can be given as

$$\mathrm{P}_{\hat{\beta}j}{}^{SCAD} = \begin{cases} (|\hat{\beta}|_j - \lambda) + sign(\hat{\beta}_j), & \text{if } |\hat{\beta}_j| < 2\lambda \\ \frac{(a-1)\hat{\beta}_j - sign(\hat{\beta}_j)a\lambda}{a-2} & \text{if } 2\lambda < |\hat{\beta}_j| \leq a\lambda \\ \hat{\beta}_j & \text{if } \hat{\beta}_j > a\lambda \end{cases}$$

This thresholding rule involves two unknown parameters λ and a. Theoretically, the best pair (λ, a) could be obtained using two dimensional grids search using some criteria like cross validation methods. However, such an implementation could be computationally expensive. Based on Bayesian statistical point of view and simulation studies, Fan and Li suggested a $=$ 3.7 is a good choice for various problems and selected by cross validation method.

3.9 Criteria for evaluating subset regression models:

In the evaluation of subset regression models, in order to get the best, we make use of the following as a measure of adequacy. We have, the coefficient of multiple determinations, adjusted coefficient of multiple determinations, the residual mean square.

3.10 Coefficient of multiple determinations R^2:

The coefficient of multiple determinations is the proportion/percentage of the total variation in the dependent variable observed that can be explained by the independent variables. It measures the strength of the relationship between the dependent and the independent variables. It is given as:

$$R^2 = 1 - \frac{SSE}{SST}$$

3.11 Adjusted R^2:

To avoid the difficulties of interpreting R^2, the use of adjusted R^2 statistics is preferable, defined for a p-term equation as

$$\bar{R}_p^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2)$$

The $\bar{R}_p^2$ statistics does not necessarily increase as additional regressors are introduced into the model, except the partial F- statistic for testing the significance of the s additional regressors exceeds one. The criterion for selection of an optimum subset model is to choose the model that has a maximum $\bar{R}_p^2$.

3.12 Residual mean square:

The residual mean square for a subset residual model may be used as a model evaluation criterion. The MSE(P) experiences an initial decrease, then it stabilizes and eventually may increase as p increases, this is because the sum of square error, SSE(p) always decreases as p increases.

$$MS_E(p) = \frac{SS_E(p)}{n-p}$$

The criterion for selection of an optimum subset model is to choose the model with the following:
a) The minimum $MS_E(p)$;
b) The value of p such that $MS_E(p)$, is approximately equal to MS_E. Note: the subset regression model that minimizes $MS_E(p)$ will also maximize $\bar{R}_p^2$.

Proof:

$$\bar{R}_p^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2)$$

$$= 1 - \left(\frac{n-1}{n-p}\right)\frac{SS_E(p)}{S_{yy}}$$

$$= 1 - \left(\frac{n-1}{s_{yy}}\right)\frac{SS_E(p)}{n-p}$$

We recall:

$$MS_E(p) = \frac{SS_E(p)}{n-p}$$

$$\bar{R}_p^2 = 1 - \frac{n-1}{s_{yy}} MS_E(p)$$

Thus, the criteria minimum residual mean square and maximum adjusted coefficient of multiple determinations are equivalent.

4 Numerical Solution

We discuss a numerical method that uses the traditional EM algorithm applied to finite mixture models with revised maximization in the M step for LASSO.

4.1 Maximization of the Penalized Log-Likelihood Function

Let $(x_1, y_1), ..., (x_n, y_n)$ be a random sample of observations from the FMR model (1). In the context of finite mixture models the EM algorithm of Dempster et al. (1977) provides a convenient approach to the optimization problem. However, due to Condition P_0 which is essential to achieve sparsity, $p_{nk}(\beta)$'s are not differentiable at $\beta = 0$. The Newton-Raphson algorithm can not be directly used in the M-step of the EM algorithm unless it is properly adopted to deal with the single non-smooth point at $\beta = 0$. We follow FAN et al. (2001) and replace $p_{nk}(\beta)$ by a local quadratic approximation

$$p_{nk}(\beta) \simeq p_{nk}(\beta_0) + \frac{p'_n(\beta_0^m)}{2\beta_0}(\beta^2 - \beta_0^2)$$

in a neighborhood of β_0. This function increases to infinite whenever $|\beta \to \inf$ which is more suitable to our application than the simple Taylor's expansion. Let $\Psi^{(m)}$ be the parameter value after the mth iteration. We replace $p_n(\Psi)$ in the penalized log-likelihood function in (2) by the following function:

$$\tilde{p_n}(\Psi;\Psi^{(m}) = \sum_{k=1}^{K} \pi_k \sum_{j=1}^{P} \{p_{nk}(\beta_{jk}^{(m)}) + \frac{p'_n(\beta_{jk}^{(m)})}{2\beta_{jk}^{(m)}}(\beta_{jk}^2 - \beta_{jk}^{(m)^2})$$

The revised EM algorithm is as follows. Let the complete log-likelihood function be

$$l_n^c(\Psi) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}[log\pi_k + log\{f(y_i;\theta_k(x_i),\phi_k)\}]$$

where $z'_{ik}s$ are indicator variables showing the component-membership of the ith observation in the FMR model and they are unobserved imaginary variables. The penalized complete log-likelihood function is then given by $\tilde{l_n^c}(\Psi) = l_n^c(\Psi) - p_n(\Psi)$. The EM algorithm maximizes $\tilde{l_n^c}\Psi$ iteratively in two steps as follows.

E-Step: Let $\Psi^{(m)}$ be the estimate of the parameters after the mth iteration. The E-step computes the conditional expectation of the function $\tilde{l_n^c}(\Psi)$ with respect to z_{ik}, given the data (x_i, y_i), and assume the current estimate $\Psi^{(m)}$ are the true parameters of the model. The conditional expectation is found to be

$$Q(\Psi;\Psi^{(m}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} log\pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} log\{f(y_i;\theta_k(x_i),\Phi_k)\} - p_n(\Psi)$$

where the weights

$$w_{ik}^{(m)} = \frac{\pi_k^m f(y_i;\theta_k^{(m)}(x_i),\phi_k^{(m)})}{\sum_{l=1}^{K} \pi_l^{(m)} f(y_i;\theta_l^{(m)}(x_i),\phi_l^{(m)})} \tag{2}$$

M-Step: The M-step on the $(m+1)th$ iteration maximizes the function $Q(\Psi;\Psi^{(m)}$ with respect to Ψ. In a usual EM-algorithm, the mixing proportions are updated by

$$\pi_k^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}^{(m)} \tag{3}$$

which maximize the leading term of $Q(\Psi;\Psi^{(m)})$. Maximizing $Q(\Psi;\Psi^{(m)})$ itself with respect to π's will be more complex. For simplicity, we use the updating scheme (5) nevertheless. It worked well in our simulations. We now consider that π_k are constant in $Q(\Psi;\Psi^{(m)})$, and maximize $Q(\Psi;\Psi^{(m)})$ with respect to other part of the parameters in Ψ. By replacing $p_n(\Psi)$ by $\tilde{p}_n(\Psi;\Psi^{(m)})$ in $Q(\Psi;\Psi^{(m)})$, the updated regression coefficients and dispersion parameters are found from Khalili and Chen (2007). Starting from an initial value $\Psi^{(0)}$, we iterate between the E and M-steps until some convergence criterion is satisfied. When the algorithm converges, the equation

$$\frac{\delta l_n(\Psi_n)}{\delta\beta_{kj}} - \tilde{p}_{nk}(\beta_{kj}) = 0 \tag{4}$$

is satisfied (approximately) for the non-zero estimate $\hat{\beta}_{kj}$. At the same time, (6) is not satisfied when the estimated value of $\beta_k j$ is zero. This fact enables us to identify zero estimates. For other issues of numerical implementation, the paper by Hunter and Li (2005) will be helpful.

4.2 Choice of the Tuning Parameters

In using LASSO penalty functions, we need to choose the sizes of some tuning parameters γ_{nk}. The current theory only provides some guidance on the order of γ_{nk} to ensure the sparsity property. In applications, the cross-validation (CV); Davies et al. (1974), or generalized cross validation (GCV); Craven and Wahba (1979), are often used for choosing tuning parameters. Following the examples of Tibshirani (1996) and Fan and Li (2001), In here we are using a process developed by Khalili and Chen (2007).

5 Simulation Study

Our simulations are based on the Normal regression model with P=10 covariates along with 200 observations with 500 repetations. The covariate x in the simulation is generated from multivariate normal with mean 0, variance 1, and correlation $Cor(x_i, x_j) = (0.5)^{|i-j|}$. We compare the performance of different variable selection methods from a number of angles. The first is the average correct and incorrect estimated zero effects in each component of the regression model. The parameter values are

Method	Sensitivity	Specificity
Backward elimination	100	83.48
LASSO	100	99.75
Adaptive LASSO	99.96	99.80
SCAD	99.96	89.88

Table 1: Sensitivity and Specificity of differents method

$\beta = (1, -1, 2, 0, 0, 0, -1.5, 0, 1, 0)$ We denote sensitivity and specificity of the following methods for simulated data in the following table.

In here we can see that penalized method has better perfomance than backward and among penalized lasso giving us better for this simulated data.

6 Real Data Analysis

6.1 Multiple regression

Regression is a statistical tool for evaluation the relationship between one or more independent variable(s) $X_1, X_2, ...X_n$ and a single continuous dependent variable Y. it is most often used when the independent variables are not controllable, that is, when collected in a sample survey or other observational studies. Multiple regression involves more than one independent variables. However, in this work the multiple regression analysis shall consist of eleven (11) independent variables. Our dataset is Red wine quality data. This dataset is obtained from the UCL Machine Learning data source. Independent variables (based on physicochemical tests) are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and Output(dependent) variable (based on sensory data) is quality (score between 0 and 10).

6.2 Testing for homoscedasticity:

Using the scatter plot of the standardized residual against the standardized predictors on all the model obtained by the various variable selection techniques to check for constant variance, we observe that from each chart obtained that only a few points less than four of the residuals vary. This indicates the presence of outliers. But since only few points vary in the entire chart obtained, we can therefore conclude that the residuals have constant variance.

6.3 Testing for autocorrelation:

Using Durbin Watson d Statistics, we observe that the Durbin Watson d statistic value for the direct search on t, forward selection method, backward elimination method, and the stepwise regression method are 1.049, 0.986, 1.203, and 0.986. since all the values obtained falls within the range 0 and 4, we accept the null hypothesis and conclude that there is no autocorrelation between residuals overtime given by each subset regressor model given by the various techniques.

6.4 Obtaining the optimal regression model for the estimation:

Here, we have seen independent variables and an optimal regression model is required.

6.5 Using direct search on t:

From above table, we observe that the regression coefficients of the wine quality, Total listing of the Volatile acidity, Chlorides, Free sulfar di oxide, Total sulfer di oxide, PH, Sulphates and Alcohol are -1.084e+00, -1.874e+00, -4.361e-03, -3.265e-03, -4.137e-01, 9.163e-01 and 2.762e-01 and the probability associated with their t values are all smaller than the preselected level of significance 0.05. This indicates that the contribution of the seven independent variables to the model is statistically significant. Therefore, removing the insignificant independent variables, we regressed the other seven variables and the model obtained is;

$$winequality = 2.197e{+}01 - 1.084e{+}00 * Volatileacidity - 1.874e{+}00 * Chlorides - 4.361e{-}03 * freesulfer - 3.265e{-}03 * total$$

	Estimate	Standard.Error	t-value	Significant
Intercept	2.197e+01	2.119e+01	1.036	0.3002.
Fixed acidity	2.499e-02	2.595e-02	0.963	0.3357.
Volatile acidity	-1.084e+00	1.211e-01	-8.948	2e-16.
citric acid	-1.826e-01	1.472e-01	-1.240	0.2150.
residual sugar	1.633e-02	1.500e-02	1.089	0.2765.
chlorides	-1.874e+00	4.193e-01	-4.470	8.37e-06.
free sulfur dioxide	4.361e-03	2.171e-03	2.009	0.0447.
total sulfur dioxide	-3.265e-03	7.287e-04	-4.480	8.00e-06.
density	-1.788e+01	2.163e+01	-0.827	0.4086.
pH	-4.137e-01	1.916e-01	-2.159	0.0310.
sulphates	9.163e-01	1.143e-01	8.014	2.13e-15.
alcohol	2.762e-01	2.648e-02	10.429	2e-16.

Table 2: Coefficients and p-value of all independent variables

Which have a R^2 value of 0.3606 which shows that 36% of the total variation can be explained by the independent variables.

6.5.1 Testing for model adequacy:

a) Hypothesis: H_0 =The model is not adequate; H_1 = The model is adequate. Using a 5% level of significance.

Model	Sum squares	DF	Mean square	F	Sig.
Regression	375.8	11	34.1636	81.35	2.2e-16
Residual	661.41	1587	0.420		
Total	1037.21	1598			

Table 3: Anova table using all variables

b) Interpretation: From the result above with a F-ratio of 81.35 which is significant at 0.000<0.05, we reject the null hypothesis statement that the model is not adequate and conclude that the model is statistically adequate.

6.5.2 Test for parameter significance:

a) Hypothesis:H_0=The parameter is not significant; H_1 = The parameter is not significant. Using a 5% level of significance.

b) Interpretation: From the result, we observed that seven independent variables and have unstandardized coefficients respectively with t-values respectively are significant at smaller than 0.05. Therefore, we reject the null hypothesis statement that the parameters are not significant and conclude that these seven parameters are statistically significant.

6.6 Using the backward elimination method:

Using backward elimination method, the appropriate model for this analysis, using the following table:

Using backward elimination process, we can get estimate and model efficacy of this model and following table shows the output of these variable,

Using backward elimination we have our model, $quality = 4.4300987 - 1.012753 * volatile.acidity - 2.017814 * chlorides + 0.005077 * free.sulfur.dioxide - 0.003482 * total.sulfur.dioxide - 0.482661 * pH + 0.882665 * sulphates + 0.289303 * alcohol$, R^2=0.3595 which shows that 35.9% of the total variation in wine quality can be explained by the seven independent variables.

6.6.1 Testing for model adequacy:

a) Hypothesis: H_0 =The model is not adequate; H_1 =The model is adequate. Using a 5% level of significance Table 5.

	DF	Sum Square	RSS	AIC
none			667.54	-1380.8.
-free sulfur dioxide	1	2.394	669.93	-1377.1
-pH	1	7.073	674.61	-1365.9
-total sulfur dioxide	1	10.787	678.32	-1357.2
-chlorides	1	10.809	678.35	-1357.1
-sulphates	1	27.060	694.60	-1319.2
-volatile acidity	1	42.318	709.85	-1284.5
-alcohol	1	124.483	792.02	-1109.4

Table 4: List of variables using Backward elimination

	Estimate	Std.Error	t-value	sig.
Intercept	4.4300987	0.4029168	10.995	2e-16
volatile acidity	-1.0127527	0.1008429	-10.043	2e-16
chlorides	-2.0178138	0.3975417	-5.076	4.31e-07
free sulfur dioxide	0.0050774	0.0021255	2.389	0.017
total sulfur dioxide	-0.0034822	0.0006868	-5.070	4.43e-07
pH	-0.4826614	0.1175581	-4.106	4.23e-05
sulphates	0.8826651	0.1099084	8.031	1.86e-15
alcohol	0.2893028	0.0167958	17.225	2e-16

Table 5: Table of last Backward elimination

b) Interpretation: From the result in table 5 above, we observe that with a F-ratio of 127.6 which is significant at 0.05, we reject the null hypothesis statement that the model is not adequate and conclude that the model is adequate.

6.6.2 Test for parameter significance:

a) Hypothesis: H_0 =The parameter is not significant; H_1 = The parameter is significant. Using a 5% level of significance, Table 4.

b) Interpretation: From the result in the table above, we observed that the independent variables volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol and have coefficients of 4.430099, -1.012753, -2.017814, 0.005077, -0.003482, -0.482661, 0.882665, 0.289303 respectively with p-values all less than 0.05 indicating that the contribution of all the regressors to the model is statistically significant.

6.7 Using the Forward selection method:

Using forward selection method, the appropriate model for this analysis, using the following table:

Using forward selection process, we can get estimate and model efficacy of this model and following table shows the output of these variable,

Using forward selection we have our model, $quality = 4.43009 - 1.012753 * volatile.acidity - 2.017814 * chlorides + 0.005077 * free.sulfur.dioxide - 0.003482 * total.sulfur.dioxide - 0.482661 * pH + 0.882665 * sulphates + 0.289303 * alcohol$, R^2=0.3595 which shows that 35.9% of the total variation in wine quality can be explained by the seven independent variables.

6.7.1 Testing for model adequacy:

a) Hypothesis: H_0 =The model is not adequate; H_1 =The model is adequate. Using a 5% level of significance Table 8.

b) Interpretation: From the result in table 7 above, we observe that with a F-ratio of 127.6 which is significant at 0.05, we reject the null hypothesis statement that the model is not adequate and conclude that the model is adequate.

Model	DF	Sum Sq	Mean Sq	F	Sig.
Regression	7	374.63	53.518	127.6	2.2e-16
Residuals	1591	667.54	0.420		
Total	1598	1042.17			

Table 6: Anova table using backward elimination

	DF	Sum Square	RSS	AIC
none			667.54	-1380.8 + citric acid
1	0.47480	667.06	-1379.9	
+ residual sugar	1	0.16673	667.37	-1379.2
+ density	1	0.03079	667.51	-1378.9
+ fixed acidity	1	0.00663	667.53	-1378.8

Table 7: List of variables using forward selection

6.7.2 Test for parameter significance:

a) Hypothesis: H_0 =The parameter is not significant; H_1 = The parameter is significant. Using a 5% level of significance, Table 7.

b) Interpretation: From the result in the table above, we observed that the independent variables volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol and have coefficients of 4.430099, -1.012753, -2.017814, 0.005077, -0.003482, -0.482661, 0.882665, 0.289303 respectively with p-values all less than 0.05 indicating that the contribution of all the regressors to the model is statistically significant.

6.8 Using the stepwise selection method:

Using stepwise selection method, the appropriate model for this analysis, using the following table:

Using stepwise selection process, we can get estimate and model efficacy of this model and following table shows the output of these variable,

Using stepwise selection we have our model, $quality = 4.43009 - 1.012753 * volatile.acidity - 2.017814 * chlorides + 0.005077 * free.sulfur.dioxide - 0.003482 * total.sulfur.dioxide - 0.482661 * pH + 0.882665 * sulphates + 0.289303 * alcohol$, R^2=0.3595 which shows that 35.9% of the total variation in wine quality can be explained by the seven independent variables.

6.8.1 Testing for model adequacy:

a) Hypothesis: H_0 =The model is not adequate; H_1 =The model is adequate. Using a 5% level of significance Table 11.

b) Interpretation: From the result in table 7 above, we observe that with a F-ratio of 127.6 which is significant at 0.05, we reject the null hypothesis statement that the model is not adequate and conclude that the model is adequate.

6.8.2 Test for parameter significance:

a) Hypothesis: H_0 =The parameter is not significant; H_1 = The parameter is significant. Using a 5% level of significance, Table 10.

b) Interpretation: From the result in the table above, we observed that the independent variables volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol and have coefficients of 4.430099, -1.012753, -2.017814, 0.005077, -0.003482, -0.482661, 0.882665, 0.289303 respectively with p-values all less than 0.05 indicating that the contribution of all the regressors to the model is statistically significant.

6.9 LASSO selection procedure:

Using LASSO we draw a plot and see that, as the constraint is relaxed, more attributes shows up. alcohol is the first to emmerge, followed by sulphates and volatile.acidity. density is the last one. We also note that all the coefficients are negative except sulphates and alcohol. Plot is given below:

	Estimate	Std.Error	t-value	sig.
Intercept	4.4300987	0.4029168	10.995	2e-16
alcohol	0.2893028	0.0167958	17.225	2e-16
volatile acidity	-1.0127527	0.1008429	-10.043	2e-16
sulphates	0.8826651	0.1099084	8.031	1.86e-15
total sulfur dioxide	-0.0034822	0.0006868	-5.070	4.43e-07
chlorides	-2.0178138	0.3975417	-5.076	4.31e-07
pH	-0.4826614	0.1175581	-4.106	4.23e-05
free sulfur dioxide	0.0050774	0.0021255	2.389	0.017

Table 8: Table of coefficients of forward selection method

Model	DF	Sum Sq	Mean Sq	F	Sig.
Regression	7	374.63	53.518	127.6	2.2e-16
Residuals	1591	667.54	0.420		
Total	1598	1042.17			

Table 9: Anova table using forward selection

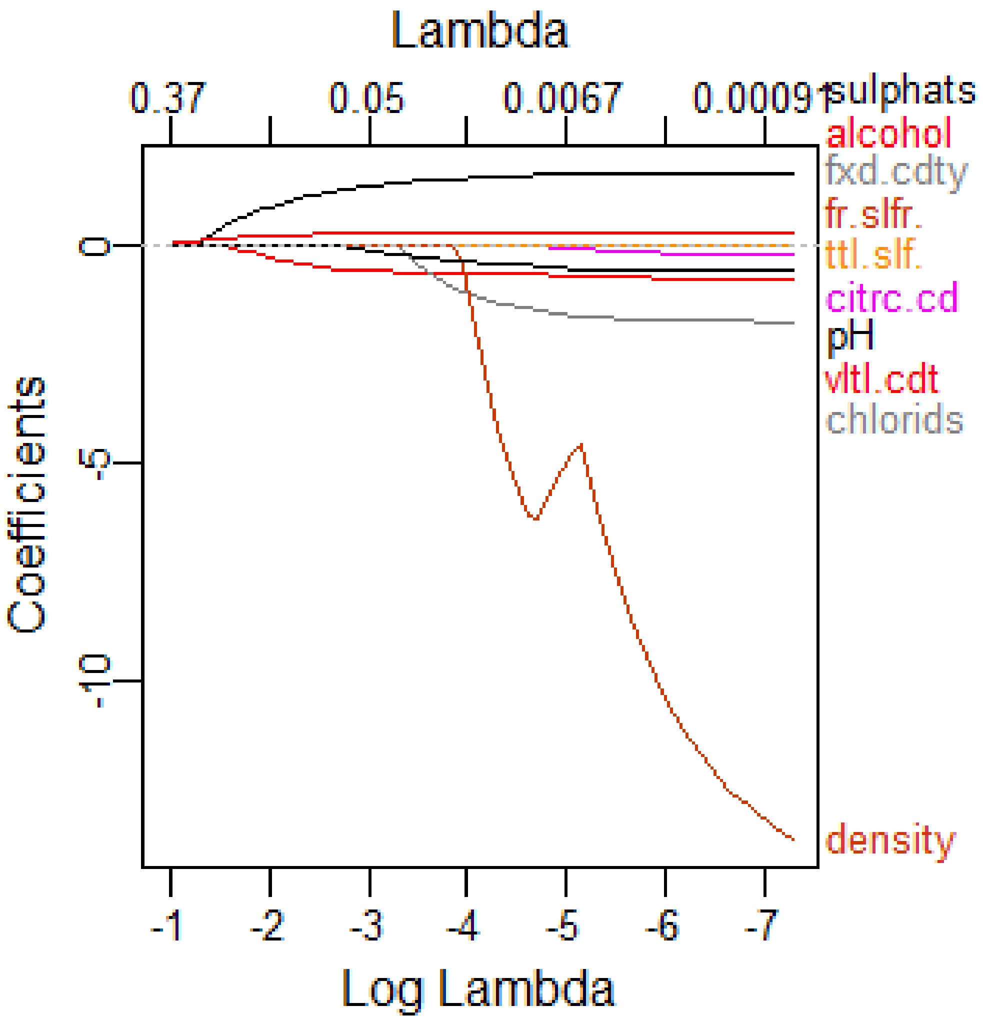

Figure 1: LASSO plot for wine quality dataset

	DF	Sum Square	RSS	AIC
none			667.54	-1380.8
+ citric acid	1	0.475	667.06	-1379.9
+ residual sugar	1	0.167	667.37	-1379.2
+ density	1	0.031	667.51	-1378.9
+ fixed acidity	1	0.007	667.53	-1378.8
- free sulfur dioxide	1	2.394	669.93	-1377.1
- pH	1	7.073	674.61	-1365.9
- total sulfur dioxide	1	10.787	678.32	-1357.2
- chlorides	1	10.809	678.35	-1357.1
- sulphates	1	27.060	694.60	-1319.2
- volatile acidity	1	42.318	709.85	-1284.5
- alcohol	1	124.483	792.02	-1109.4

Table 10: List of variables using stepwise selection

	Estimate	Std.Error	t-value	sig.
Intercept	4.4300987	0.4029168	10.995	2e-16
alcohol	0.2893028	0.0167958	17.225	2e-16
volatile acidity	-1.0127527	0.1008429	-10.043	2e-16
sulphates	0.8826651	0.1099084	8.031	1.86e-15
total sulfur dioxide	-0.0034822	0.0006868	-5.070	4.43e-07
chlorides	-2.0178138	0.3975417	-5.076	4.31e-07
pH	-0.4826614	0.1175581	-4.106	4.23e-05
free sulfur dioxide	0.0050774	0.0021255	2.389	0.017

Table 11: Table of coefficients of stepwise selection method

Now, from cross validation plot using LASSO, we can say about MSE.

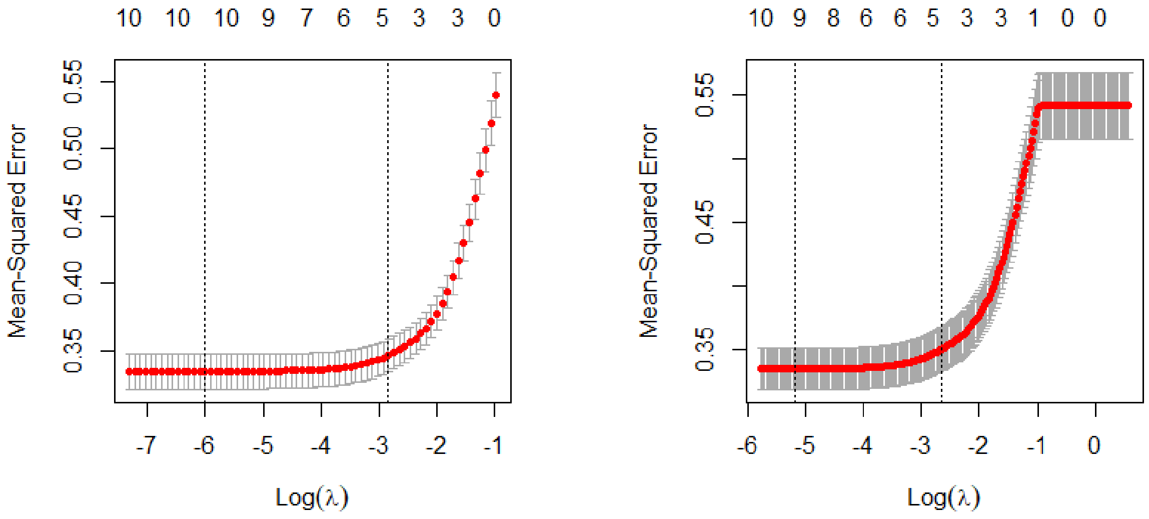

Figure 2: Cross validation plot for LASSO

This plot of cross validation shows us that as the model complexity decrease, the MSE increase. However, as there is at least 5 attributes in the model, the MSE stay low. Using LASSO selection process, we can get estimate and model efficacy of this model and following table shows the output of this procedure,

Model by lasso includes 5 variables: alcohol, sulphates, volatile.acidity, total.sulfur.dioxide, and pH.

Model	DF	Sum Sq	Mean Sq	F	Sig.
Regression	7	374.63	53.518	127.6	2.2e-16
Residuals	1591	667.54	0.420		
Total	1598	1042.17			

Table 12: Anova table using stepwise selection

	Estimate
Intercept	2.5425301221
fixed.acidity	.
volatile.acidity	-0.5659572211
citric.acid	.
residual.sugar	.
chlorides	.
free.sulfur.dioxide	.
total.sulfur.dioxide	-0.0006907149
density	.
pH	-0.0045820258
sulphates	1.2605890598
alcohol	0.2555687705

Table 13: Table of coefficients of LASSO selection method

6.10 Adaptive LASSO selection procedure:

Using Adaptive LASSO we draw a plot and see that, as the constraint is relaxed, more attributes shows up. citrc.cd is the first to emmerge, followed by sulphates and volatile.acidity. density is the last one. We also note that all the coefficients are negative except sulphates and alcohol. Plot is given below:

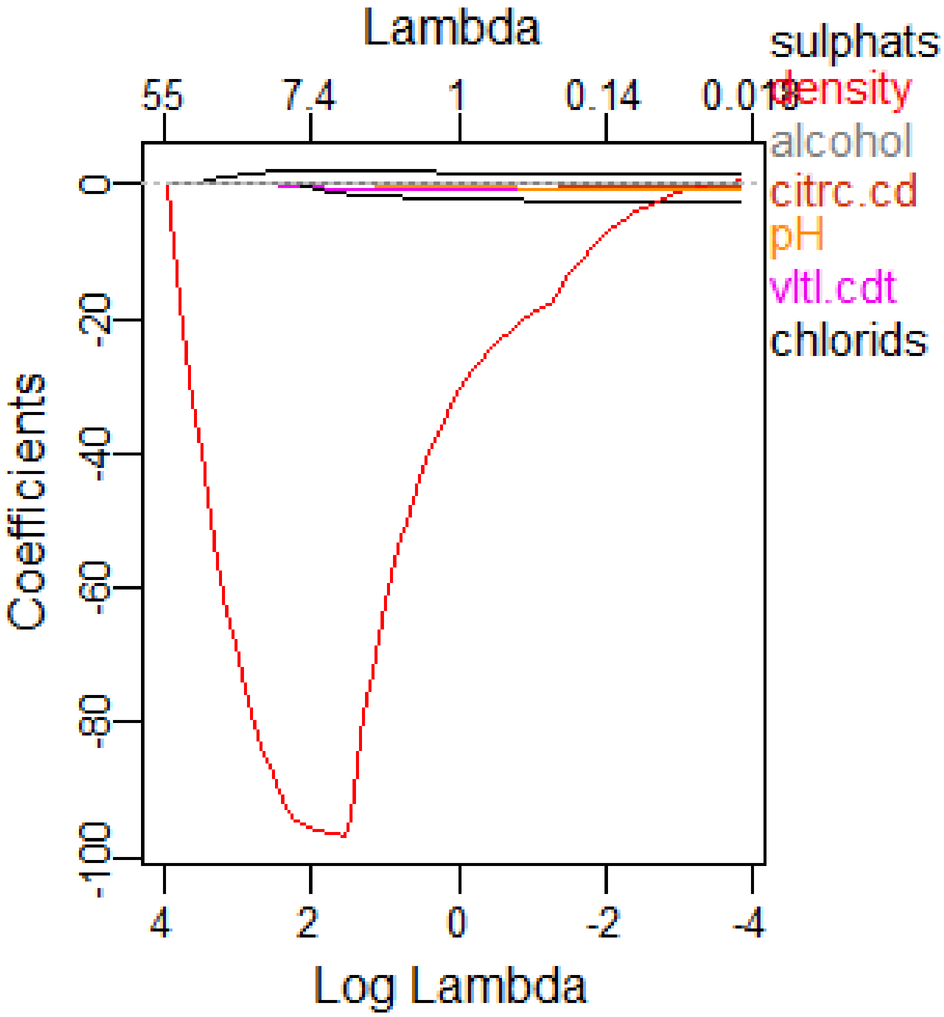

Figure 3: Adaptive LASSO plot for wine quality dataset

Now, from cross validation plot using LASSO, we can say about MSE.

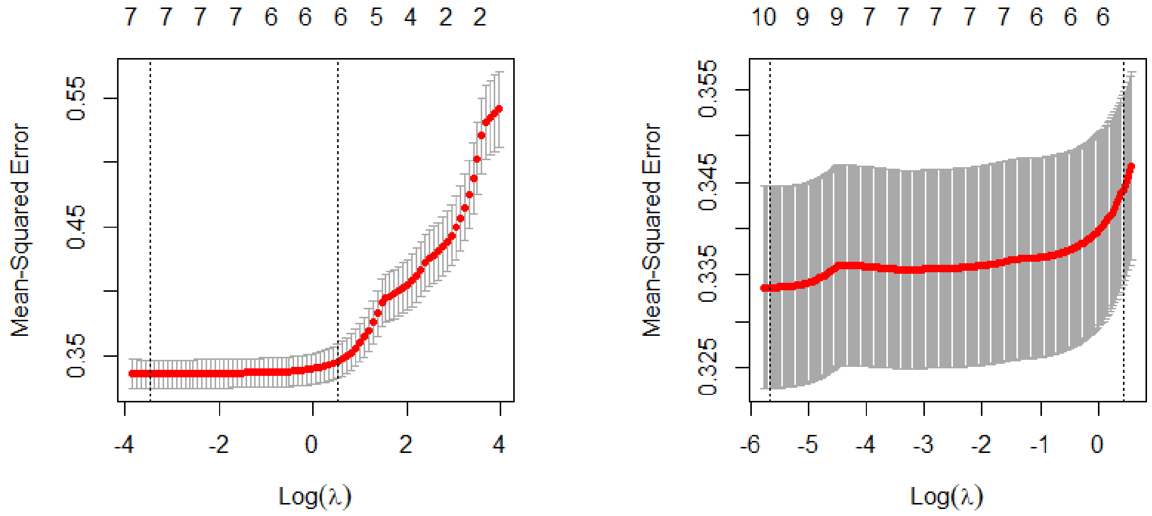

Figure 4: Cross validation plot for Adptive LASSO

This plot of cross validation shows us that as the model complexity decrease, the MSE increase. However, as there is at least 5 attributes in the model, the MSE stay low.

	Estimate
Intercept	47.7036339
fixed.acidity	.
volatile.acidity	-0.6781777
citric.acid	.
residual.sugar	.
chlorides	-1.9943453
free.sulfur.dioxide	.
total.sulfur.dioxide	.
density	-43.7719404
pH	-0.2959543
sulphates	1.8454135
alcohol	0.1827492

Table 14: Table of coefficients of Adaptive LASSO selection method

Model by Adaptive lasso includes 6 variables: alcohol, sulphates, volatile.acidity, total.sulfur.dioxide, density and pH.

6.11 SCAD procedure for variable selection:

Using SCAD we draw a plot and notice that variables enter the model one at a time, and that at any given value of , several coefficients are zero. Plot is given below:

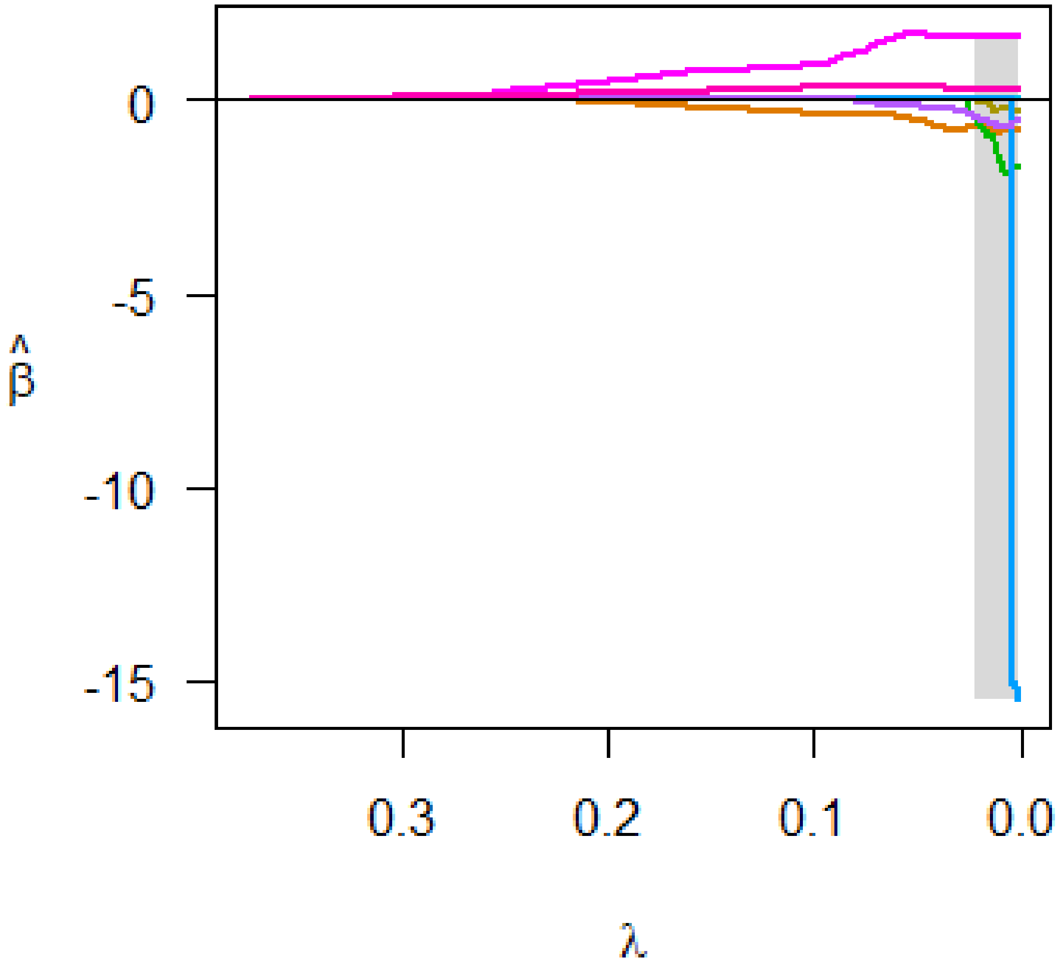

Figure 5: SCAD plot for wine quality dataset

Now, from cross validation plot using SCAD, we can say about MSE.

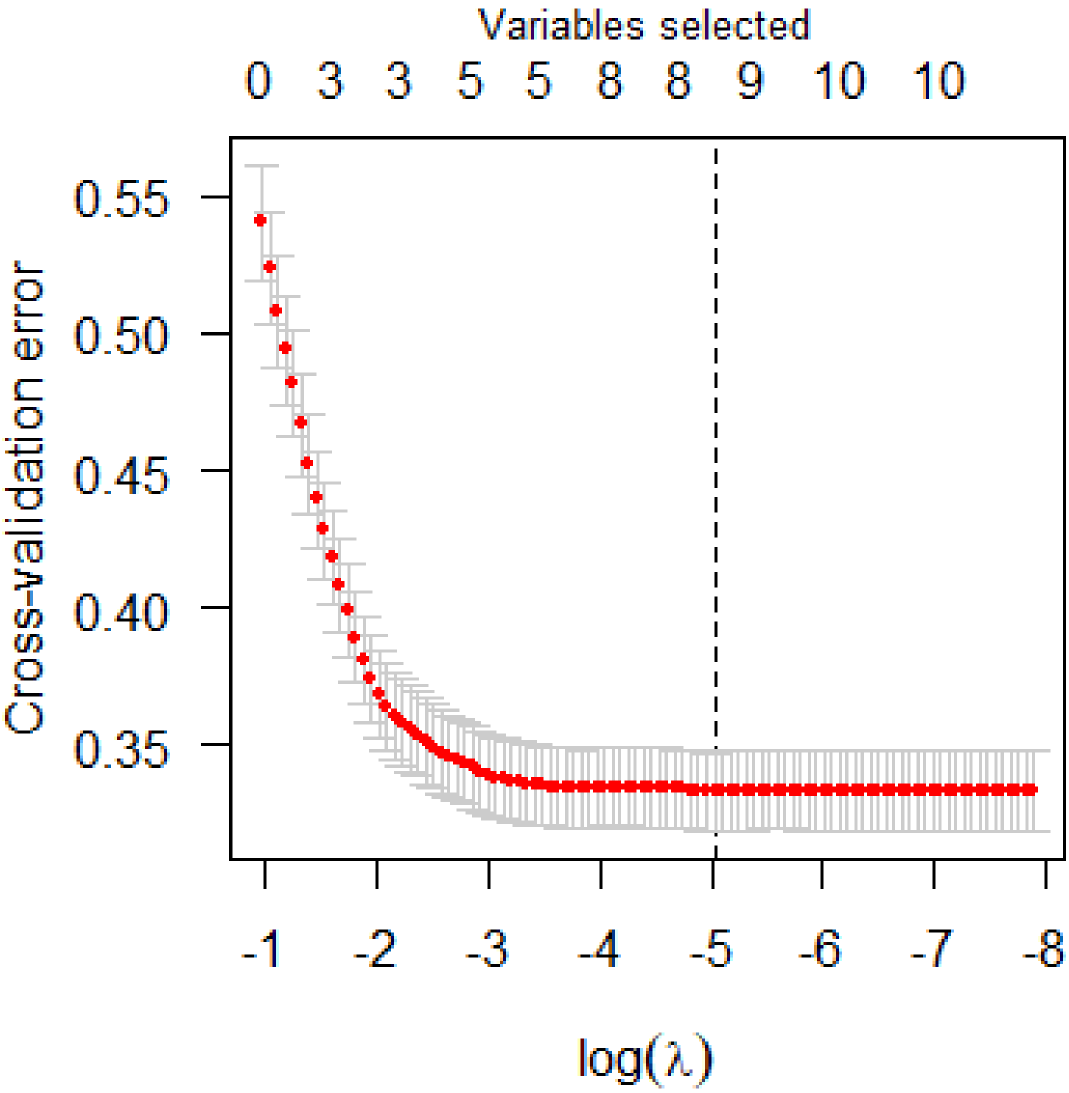

Figure 6: Cross validation plot for SCAD

	Estimate
Intercept	4.520564759
fixed.acidity	0.000000000
volatile.acidity	-0.797830137
citric.acid	-0.230647958
residual.sugar	0.000000000
chlorides	-1.939867480
free.sulfur.dioxide	0.003470711
total.sulfur.dioxide	-0.003030461
density	0.000000000
pH	-0.679705167
sulphates	1.646001566
alcohol	0.294163297

Table 15: Table of coefficients of SCAD selection method

This plot of cross validation shows us that as the model complexity decrease, the MSE increase. However, as there is at least 5 attributes in the model, the MSE stay low. Model by SCAD includes 8 variables: alcohol, sulphates, pH, total.sulfur.dioxide, free.sulfer.dioxide, Chlorides, citric.acid and volatile.acidity.

7 Conclusion

From analysis using backward elimination, stepwise, forward selection method along with the penalized likelihood approach(LASSO, Adaptive LASSO, SCAD) for variable selection in the context of regression. The penalty function is designed to be dependent on the size of the regression coefficients. The penalized procedure is shown to be consistent in selecting the most parsimonious model. It also reported a data adaptive method for selecting the tuning parameters and demonstrate the usage by extensive simulations. The penalized method with the LASSO, Adaptive LASSO and SCAD penalty functions performed as well as the conventional methods while it is computationally much more efficient. In addition, as in the example of wine quality application, the penalized method can also be used to suggest a set of plausible models to be examined by the BIC method if desired. Among the penalized method LASSO procedure works better smoothly than Adaptive LASSo and SCAD. This helps to reduce the computational burden of using the substantially other conventional methods.

References

Craven, P. and Wahba, G. (1979). Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403.

Dale, B. E. (2009). Biofuels, indirect land use change & greenhouse gas emissions: Some unexplored variables (and a call to action!!).

Davies, J., Game, C., Green, M., and Stone, F. (1974). J. chem. soc., dalton trans.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

FAN, X.-L., SUN, X.-Q., HAN, Y.-L., and ZHANG, J.-S. (2001). Theoretical analysis for the double differential cross section of reaction n+ 16 o. *Chinese Physics C*, 25(9):859–864.

Fan, Y., Huang, W., Huang, G., Li, Z., Li, Y., Wang, X., Cheng, G., and Jin, L. (2015). A stepwise-cluster forecasting approach for monthly streamflows based on climate teleconnections. *Stochastic environmental research and risk assessment*, 29(6):1557–1569.

Gunter, L., Zhu, J., and Murphy, S. (2011). Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. *Journal of biopharmaceutical statistics*, 21(6):1063–1078.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Hunter, D. R. and Li, R. (2005). Variable selection using mm algorithms. *Annals of statistics*, 33(4):1617.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038.

Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pages 249–256. Elsevier.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348.

Raftery, A. E. (1999). Bayes factors and bic: Comment on "a critique of the bayesian information criterion for model selection". *Sociological Methods & Research*, 27(3):411–427.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Wang, P., Puterman, M. L., Cockburn, I., and Le, N. (1996). Mixed poisson regression models with covariate dependent rates. *Biometrics*, pages 381–400.

Wedel, M. and Kamakura, W. A. (2012). *Market segmentation: Conceptual and methodological foundations*, volume 8. Springer Science & Business Media.

Yi, N., George, V., and Allison, D. B. (2003). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics*, 164(3):1129–1138.

Zhou, G.-X., Zhu, X.-J., Ding, X.-L., Zhang, H., Chen, J.-P., Qiang, H., Zhang, H.-F., and Wei, Q. (2010). Protective effects of mcp-1 inhibitor on a rat model of severe acute pancreatitis. *Hepatobiliary & pancreatic diseases international: HBPD INT*, 9(2):201–207.

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