Variable Selection in Finite Mixture of Regression Models

May 9, 2020

Faruk Hossain
Department of Mathematical Sciences, UNLV

Abstract

Finite mixture regression (FMR) models are frequently used in statistical modeling, often with many covariates with low significance. Variable selection techniques can be employed to identify the covariates with little influence on the response. The problem of variable selection in FMR models is studied here. Existing methods such as the Akaike information criterion (AIC) Sakamoto et al. (1986) and the Bayes information criterion (BIC) Raftery (1999) are computatinally intensive as the number of covariates and components in the mixture model increases. For these reasons Khalili and Chen (2007) introduced a Penalized likelihood-based approach for variable selection in FMR. This new method introduces penalties that depend on the size of the regression coefficients and the mixture structure.Mimicking this method, Now I am showing that the proposed estimator has the variable selection consistency and oracle property. A data Adaptive method for selecting tuning parameters and an EM-algorithm for numerical computations are shown. We conduct simulation studies to demonstrate that the proposed method performs well in comparison with existing methods. We also analyze a real data set to further illustrate the usefulness of the proposed method.

Keywords: Mixture model; EM-algorithm; 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. The methods proposed for the problem of variable selection can be roughly classified into two categories. 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.

Finite mixture models provide a flexible tool for modeling data that arise from a heterogeneous population. 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 a new variable selection procedure proposed by Khalili and Chen (2007) for FMR models based on these methods. They propose new class of penalty functions to be used for variable selection in FMR models. They investigate methods for selecting tuning parameters adaptively and develop an EM algorithm for numerical computations. The new method for variable selection is shown to be consistent. The performance of the method is studied theoretically and by simulations. Our simulations indicate that the new 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 finite mixture models along with their identifiability. Section 3 illustrates the penalized likelihood-based approach for variable selection with FMR models. 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 Finite Mixture Of Regression Models

Finite mixture models are a powerful class of models for dealing with heterogeneity in a sample of observations. This is achieved by clustering observations into a small number of sub-populations or mixture components, and using a statistical model to define the characteristics of each component McLachlan and Peel (2004), Frühwirth-Schnatter (2006). Let Y be a response variable of interest and let $x = (x_1, x_2, ..., x_p)^\intercal$ be the vector of covariates believed to have an effect on Y. The FMR model is defined as follows.

Definition 1 Let $G = f(y; \theta, \phi); (\theta, \phi) \in \Theta \times (0, \infty)$ be a family of parametric density functions of Y with respect to a σ-finite measure , where $\Theta \subset R$ and Φ is a dispersion parameter. We say that (x,Y) follows a FMR model of order K if the conditional density function of Y given x has the form

$$f(y; x, \Psi) = \sum_{k=1}^{K} \pi_k f(y; \theta_k(x), \phi_k) \tag{1}$$

with $\theta_k(x) = h(x^\intercal \beta_k), k = 1, 2, ..., K$, for a given link function h(·), and for some $\Psi = (\beta_1, \beta_2, ..., \beta_K, \phi, \pi)$ with $\beta_k = (\beta_{k1}, \beta_{k2}, ..., \beta_{kP})^\intercal, \phi = (\phi_1, \phi_2, ..., \phi_K)^\intercal$, $\pi = (\pi_1, \pi_2, ..., \pi_{K-1})^\intercal$ such that $\pi_k > 0, \pi_k \in [0; 1]$ and $\sum_{k=1}^{K} \pi_k = 1$.

Model (1) can be generalized to allow the π_k values to be functions of x. Here restrict ourselves to the current model. The density function $f(y; \theta, \phi)$ can take many parametric forms, including binomial, normal, and Poisson. In some FMR models, the dispersion parameters, ϕ_k's are assumed to be equal. FMR models combine the characteristics of regression models with those of finite mixture models. Like any regression model, the FMR model is used to study the relationship between response variables and a set of covariates. At the same time, the conditional distribution of the response variable Y given the covariates is a finite mixture.

A potential problem associated with finite mixture models is their identifiability, which is the basis for any meaningful statistical analysis. In some classes of finite mixture models, a single density function can have representations corresponding to different sets of parameter values. When no two sets of parameter values specify the same distribution, the model is identifiable. Many finite mixture models, including mixtures of binomial, multinomial, normal, and Poisson distributions, are identifiable under some conditions see (Titterington, Smith, and Markov 1985).

Definition 2 Consider a FMR model with the conditional density function given in (1). For a given design matrix $(x_1, x_2, ..., x_n)^\intercal$, the FMR model is said to be identifiable if for any two parameters Ψ and Ψ^*,

$$\textstyle\sum_{k=1}^{K} \pi_k f(y; \theta_k(x_i), \phi_k) = \sum_{k=1}^{K^*} \pi_k^* f(y; \theta_k^*(x_i), \phi_k^*)$$

for each i = 1, . . . , n and all possible values of y, implies that $K = K^*$ and $\Psi = \Psi^*$. When we exchange the order of two regression components, the parameter Ψ changes. In the foregoing definition, we interpret $\Psi = \Psi^*$ up to a permutation. The identifiability of an FMR model depends on several factors, such as component densities $f(y; \theta, \phi)$, the maximum possible order K, and the design matrix $(x_1, x_2, ..., x_n)^\intercal$. Hennig (2000) observed that for fixed designs, a sufficient condition for identifiability is that the design points do not fall in the union of any K linear subspaces of $(P - 1)$ dimension, in addition to some usual conditions on the component density. This condition is applicable to Poisson and normal FMR models. If the x_i values are also a random sample from a marginal density $f(x)$ that does not depend on Ψ, then $f(x)$ must not have all of its mass in up to K of $(P - 1)$ dimensional linear subspaces. Some discussion has been provided by Wang et al. (1996). In this article we assume that the FMR model under consideration is identifiable with the given or random design. These two illustrative examples below give rise to the general notion of a mixture model which assumes each observation is generated from one of K mixture components.

Before moving on, we make one small pedagogical note that sometimes it makes us confuse to mixture models. We might recall that if X and Y are independent normal random variables, then $Z = X + Y$ is also a normally distributed random variable. From this, we might wonder why the mixture models above aren't normal. The reason is that $X + Y$ is not a bivariate mixture of normals. It is a linear combination of normals. A random variable sampled from a simple Gaussian mixture model can be thought of as a two stage process. First, we randomly sample a component, then we sample our observation from the normal distribution corresponding to that component. This is clearly different than sampling X and Y from different normal distributions, then adding them together.

In below we add two graphs first one is for Galaxy data set from MASS packages. Where a numeric vector of velocities in km/sec of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. we plot Density against Velocity. And second on is patient dataset where we plot density against number of patients. From graph we clearly see that first one has at least 3 components and second one has 2 components and each of them follow gaussian. So they are mixture of gaussian models.

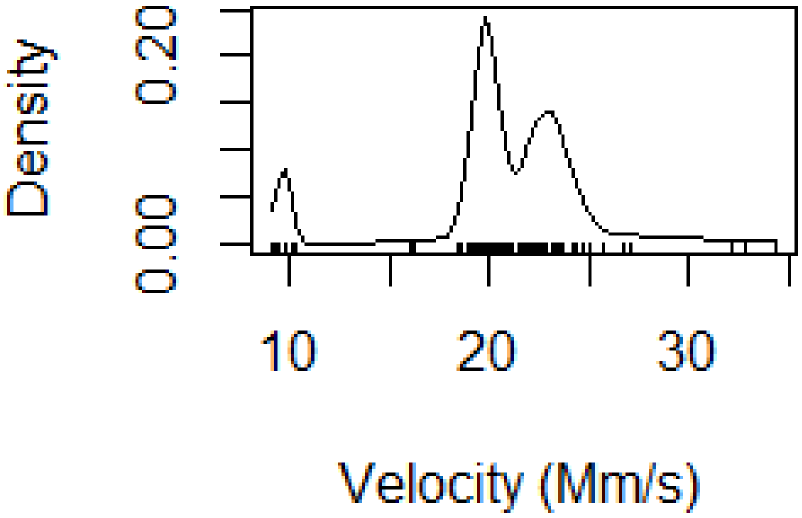

Figure 1: Plot of galaxy dataset

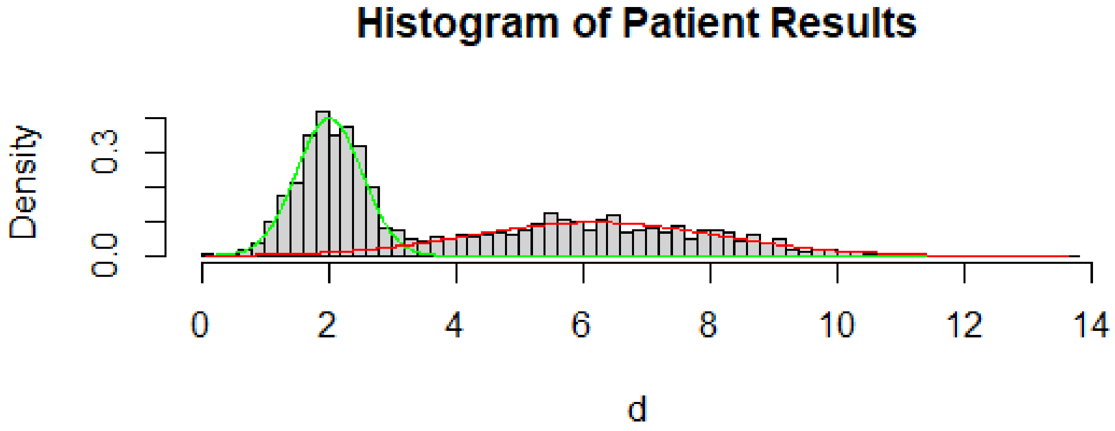

Figure 2: Plot of patients dataset

3 The Method For Variable Selection

In statistics, variable selection, also known as attribute selection or variable subset selection, feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons: First one like simplification of models to make them easier to interpret by researchers/users. Secondly, shorter training times, to avoid the curse of dimensionality, enhanced generalization by reducing overfitting(reduction of variance).

In the case where x is random, we assume that its density $f(x)$ is functionally independent of the parameters in the FMR model. Thus the statistical inference can be done based purely on the conditional density function specified in Definition 1.

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a sample of observations from the FMR model (1). The (conditional) log-likelihood function of Ψ is given by

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

When the effect of a component of x is not significant, the corresponding ordinary maximum likelihood estimate is often close to, but not equal to 0. Thus this covariate is not excluded from the model. To avoid this problem, we may study submodels with various components of x excluded, as is done by AIC and BIC. However, the computational burden of these approaches is heavy and should be avoided. The approach that we consider here is as follows. We define a penalized log-likelihood function as

$$\widetilde{l}_n = l_n(\Psi) - p_n(\Psi) \tag{2}$$

With penalty function

$$p_n(\Psi) = \sum_{k=1}^{K} \pi_k \{ \sum_{j=1}^{P} p_{nk}(\beta_{kj}) \} \tag{3}$$

where the $p_{nk}(\beta_{kj})$ values are nonnegative and nondecreasing functions in $|\beta_{kj}|$. By maximizing $\widetilde{l}_n(\Psi)$ that contains a penalty, there is a positive chance of having some estimated values of β equaling 0 and thus of automatically selecting a submodel. Thus the procedure combines the variable selection and parameter estimation into one step and reduces the computational burden substantially. In (3) we choose the penalty imposed on the regression coefficients within the kth component of the FMR model to be proportional to π_k. This is in line with the common practice of relating the penalty to the sample size. The virtual sample size from the kth subpopulation is proportional to π_k, and this choice enhances the power of the method in our simulations.

When some prior information is available on the importance of a covariate's effects within the components of the FMR model, covariate-specific penalty functions may be used. In general, we should choose appropriate penalty functions to suit the need of the application, under the guidance of statistical theory. The following three penalty functions have been investigated in the literature in a number of contexts, and we use them to illustrate the theory that developed By Khalili and Chen (2007) for the FMR models:

- L_1-norm penalty: $p_{nk}(\beta)$=$\gamma_{nk}\sqrt{n}|\beta|$
- HARD penalty: $p_{nk}(\beta)$=$\gamma_{nk}^2 - (\sqrt{n}|\beta| - \gamma_{nk})^2 I(\sqrt{n} \times |\beta| < \gamma_{nk})$
- SCAD penalty: Let $(.)_+$ be the positive part of a quantity, $p'_{nk}(\beta)$=$\gamma_{nk}\sqrt{n} I\{\sqrt{n}|\beta| \leq \gamma_{nk}\}$+$\frac{\sqrt{n}(a\gamma_{nk} - \sqrt{n}|\beta|)_+}{a-1} I\{\sqrt{n}|\beta| > \gamma_{nk}\}$.

The L_1-norm penalty was used in LASSO by Tibshirani (1996); the other two have been discussed by FAN et al. (2001). The constants $\gamma_{nk} > 0$ and $a > 2$ are chosen based on how strenuously the procedure tries to eliminate the covariates from the model. In applications, these choices may be based on some prior information; that is, the constants may be chosen subjectively by the data analysts or by a data-driven method. We call the penalty functions $p_n(.)$ in (3) constructed from LASSO, HARD, and SCAD the MIXLASSO, MIXHARD, and MIXSCAD penalties.

The three penalty functions have similar properties with some subtle differences. Maximizing the penalized likelihood is equivalent to constrained maximization. The penalty function of LASSO is convex and thus advantageous for numerical computation. It tends to reduce all effects by similar amounts until the estimated effect is reduced to 0. When the penalty is increased, SCAD reduces smaller effects faster than larger effects. Intuitively, HARD should work more like SCAD, although less smoothly.

4 Asymptotic Properties

The regression coefficient vector β_k in the kth component into $\beta_k^\intercal = \{\beta_{1K}^\intercal, \beta_{2K}^\intercal\}$ such that β_{2k} contains the 0 effects. In general, the set of non-zero effects β_{1k} may depend on k. We choose not to use more complex notation to reflect this fact without loss of generality. Naturally, we split the parameter $\Psi^\intercal = (\Psi_1^\intercal, \Psi_2^\intercal)$ such that $\Psi_2^\intercal$ contains all zero effects, namely $\beta_{2k} : k = 1, ..., K$. The vector of true parameters is denoted as Ψ_0. The components of Ψ_0 are denoted with a superscript such as β^0_{kj}.

Our asymptotic results are presented with the help of the quantities:

$$a_n = \underset{k,j}{max}\{p_{nk}(\beta^0_{kj})/\sqrt{n} : \beta^0_{kj} \neq 0\},$$
$$b_n = \underset{k,j}{max}\{p'_{nk}(\beta^0_{kj})/\sqrt{n} : \beta^0_{kj} \neq 0\},$$
$$c_n = \underset{k,j}{max}\{p''_{nk}(\beta^0_{kj})/\sqrt{n} : \beta^0_{kj} \neq 0\}$$

where $p'_{nk}(\beta)$ and $p''_{nk}(\beta)$ are the first and second derivatives of the function $p_{nk}(\beta)$ with respect to β. The asymptotic results will be based on the following conditions on the penalty functions $p_{nk}(.)$:

P_0. For all n and k, $p_{nk}(0) = 0$, and $p_{nk}(\beta)$ is symmetric and nonnegative. In addition, it is nondecreasing and twice differentiable for all *beta* in (0,∞) with at most a few exceptions.

P_1. As n→∞, $a_n = o(1 + b_n)$, and $c_n = o(1)$.

P_2. For $N_n = \{\beta; 0 < \beta \leq n^{-1/2} log(n)\}$, $\underset{n\to\inf}{lim}\ \underset{\beta\in N_n}{inf}\ \{p'_{nk}{}_{nk}(\beta)/\sqrt{n} = \inf$

Conditions P_0 and P_2 are needed for sparsity—namely, consistent variable selection. Condition P_1 is used to preserve the asymptotic properties of the estimators of nonzero effects in the model. To develop the asymptotic theory, some commonly used regularity conditions are needed on the joint density function $f(z; \Psi)$ of $Z = (x, Y)$. We find this from Khalili and Chen (2007). Some theorem which are needed to show that FMR followed regularity conditioned and asymptotic properties are shown by Khalili and Chen (2007).

5 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.

5.1 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

$$p_{nk}(\beta) \simeq p_{nk}(\beta_0) + \tfrac{p'_n(\beta^m_0)}{2\beta_0}(\beta^2 - \beta^2_0)$$

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:

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

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

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

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 $\tilde{l^c_n}(\Psi) = l^c_n(\Psi) - p_n(\Psi)$. The EM algorithm maximizes $\tilde{l^c_n}\Psi$ 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 $\tilde{l^c_n}(\Psi)$ 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

$$w_{ik}^{(m)} = \frac{\pi_k^m f(y_i;\theta_k^{(m)}(x_i),\phi_k^{(m)})}{\sum_{l=1}^{K}\pi_l^{(m)} f(y_i;\theta_l^{(m)}(x_i),\phi_l^{(m)})} \tag{4}$$

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

$$\pi_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{ik}^{(m)} \tag{5}$$

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

$$\frac{\delta l_n(\Psi_n)}{\delta\beta_{kj}} - \tilde{p}_{nk}(\beta_{kj}) = 0 \tag{6}$$

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.

5.2 Choice of the Tuning Parameters

In using MIXLASSO, MIXHARD, MIXSCAD and other 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).

6 Simulation Study

Our simulations are based on the Normal FMR model $\pi N(x^{\intercal}\beta_1,\sigma^2) + (1-\pi)N(x^{\intercal}\beta_2,\sigma^2)$ with $\sigma^2 = 1$ and P = 5. We assume K = 2 is known. When K is unknown, one may use BIC to select K under the full regression model. When $\pi = 0.5$ we found that $\hat{K} = 2$ in 996 simulations out of 1000. When $\pi = 0.1$, the data do not contain enough information to choose K consistently.

The covariate x in the simulation is generated from multivariate normal with mean 0, variance 1, and correlation $Cor(x_i, x_j) = (0.5)^{|i-j|}$. Table 1 specifies the regression coefficients β_1, β_2 and three choices of mixing proportion π. The M_1 and M_2 represent the FMR models with parameter values given in the table.

Table 1: Regression coefficients in the Normal Finite Mixture of Regression Model

Parameters	M_1	M_2
β_1	(1, 0, 0, 3, 0)	(1, 0.6, 0, 3, 0)
β_2	(-1, 2, 0, 0, 3)	(-1, 0, 0, 4, 0.7)
π	0.5, 0.3, 0.1	0.5, 0.3, 0.1

Now from the simulated data we are finding the Maximum likelihood estimate for normal mixture using fmrs package in r. As well we are reporting Penalized Maximum Likelihood Estimation for variable selection (MixLasso, MixScad and MixHard) and compare them from the table.

Table 2: Maximum Likelihood Estimate of Finite Mixture of Regression Model ($\pi = 0.5$)

	Comp.1	Comp.2
Intercept	-0.406	-0.005
X.1	1.622	1.004
X.2	-0.094	-0.000
X.3	0.870	1.077
X.4	1.835	1.772
X.5	-0.340	-0.175
X.6	-0.132	0.572
X.7	-0.319	0.308
X.8	3.594	3.409
X.9	0.555	0.600
X.10	-1.552	-0.647

Table 3: Penalized Maximum Likelihood Estimate of Finite Mixture of Regression Model(MIXLASSO)

	Comp.1	Comp.2
X.1	1.996	0.000
X.2	-0.010	-0.074
X.3	0.000	3.054
X.4	2.906	-0.052
X.5	-0.921	1.128
X.6	0.001	0.424
X.7	0.000	0.000
X.8	3.952	2.874
X.9	0.623	0.084
X.10	-1.968	1.067

Table 4: Penalized Maximum Likelihood Estimate of Finite Mixture of Regression Model(MIXSCAD)

	Comp.1	Comp.2
X.1	2.007	-0.000
X.2	-0.000	-0.000
X.3	-0.000	3.024
X.4	2.948	-0.000
X.5	-0.966	1.257
X.6	0.000	0.095
X.7	0.000	0.000
X.8	3.971	2.995
X.9	0.663	0.000
X.10	-2.015	1.114

Table 5: Penalized Maximum Likelihood Estimate of Finite Mixture of Regression Model(MIXHARD)

	Comp.1	Comp.2
X.1	2.037	0.000
X.2	-0.071	-0.095
X.3	0.024	3.085
X.4	2.950	-0.085
X.5	-0.987	1.146
X.6	0.045	0.426
X.7	-0.013	-0.000
X.8	3.967	2.885
X.9	0.661	0.085
X.10	-2.012	1.074

From the simulated table we can see that the three penalty functions have similar properties with some subtle differences. When the constraint is tightened SCAD quickly removes variables with smaller effects and leaves larger effects untouched while LASSO reduces all effects at the same rate.Again HARD is working like SCAD except less smoothly. Here we are reporting only $\pi = 0.5$ if π is small then all three provide poor estimates but for larger all perform well. For that Case MIXLASSO also provides estimate like MIXSCAD and MIXHARD. This simulation have good performances for MIXSCAD and MIXHARD in all cases.

7 Real Data Analysis

We analyzed two real data set to further demonstrate the use of the method developed by Khalili and Chen (2007). The FMR models have often been used in target analysis. The concept of target analysis is an essential element in both effectiveness and practice. According to this concept, a heterogeneous program can be divided into a number of smaller homogeneous programs, in response to differing preferences of participants. The FMR models provide a model-based approach for this data. We analyzed these two data by fitting a multi-nomial logit FMR model. The variable selection problem presents itself naturally.

7.1 Real Dataset 1

Our first dataset is breast cancer data. This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. It has 30 covariates and outcome variable is diagnosis.

7.1.1 Data Visualization

Here are some visualizations that will help us to get started to see real results.

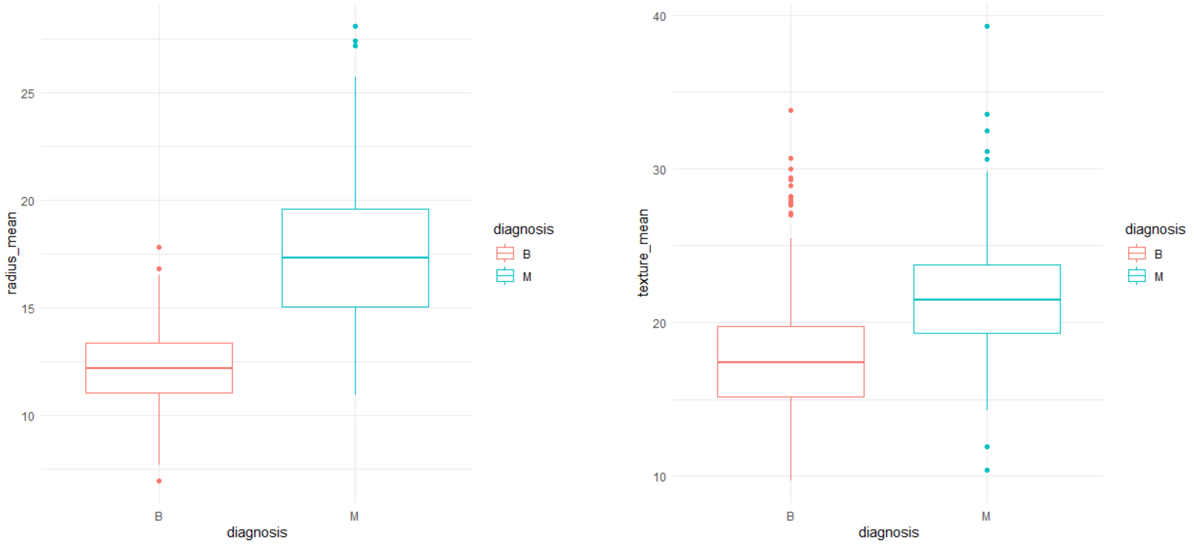

Figure 3: Boxplot of diagnosis vs radius mean and texture mean

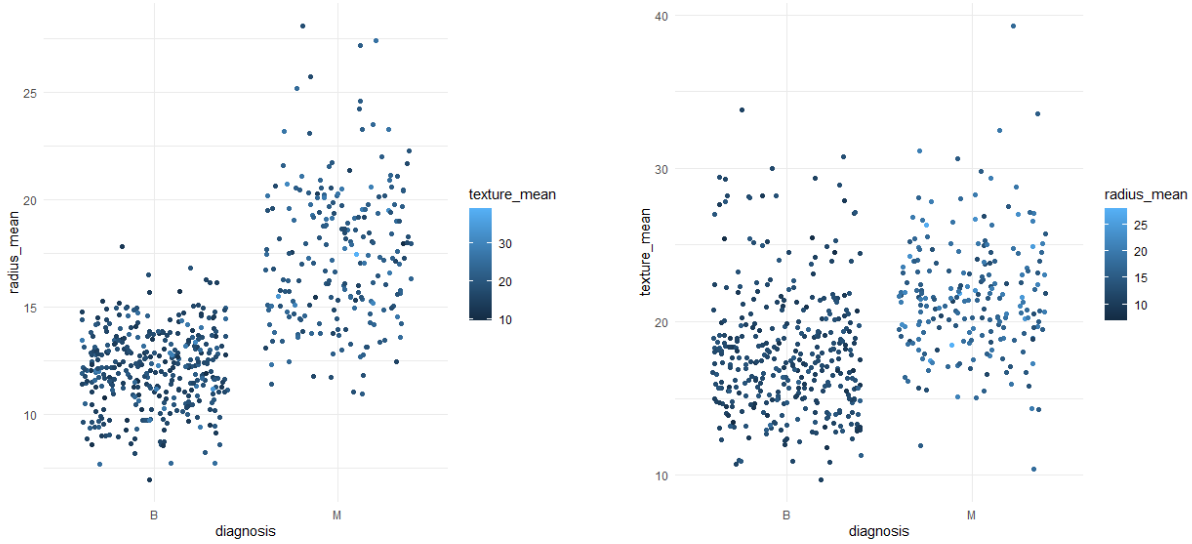

Figure 4: Plot of diagnosis vs radius mean and texture mean

From the graphs we can say, among patients with a breast cancer, When the radius mean is going to be over 25. On average, the cancer rate is lower.

7.1.2 Model and Data Analysis

We fitted a multi-nomial logit FMR model addressed by Skrondal and Rabe-Hesketh (2004), with K = 2, corresponding to two segments, to the data arise from the program type conjoint analysis. Mathematically, the FMR model is given by

$$P(y_i) = P(Y_i = y_i) = (1 - \pi)P_1(y_i) + \pi P_2(y_i)$$

Where

$$P_k(y_i) = \prod_{j=1}^{13} \prod_{a=1}^{2} [\frac{exp\{x_a^{\tau}\beta_k\}}{\sum_{l=1}^{3} exp\{x_l^{\tau}\beta_k\}}], k = 1, 2; a = 1, 2$$

The covariate x_a^{τ} is an 30×1 vector of dummy variables, corresponding to the ten attributes. Since the value of covariates x_a^{τ}'s did not change with subjects often enough, to make the parameters identifiable, an intercept term in the linear predictor can be excluded but we included it here.

we are using "fmrs" r package for finding Mixlasso, Mixscad and Mixhard for variable selection and compare them. These are penalized method. The goal is very simple: we will penalize our estimators so that the coefficients will be close (Ridge) to zero or equal (Lasso). We can reduce the complexity of the model and have more efficient estimators. It is also a great way to select the most important predictors. First we are finding maximize likelihood estimator of this data.

Table 6: Maximum Likelihood Estimate of Finite Mixture of Regression Model for breast cancer data

	Comp.1	Comp.2
Intercept	-1.022	-1.022
ml$radius_mean	-0.218	-0.218
ml$texture_mean	0.005	0.005
ml$perimeter_mean	0.024	0.024
ml$area_mean	0.000	0.000
ml$smoothness_mean	0.085	0.085
ml$compactness_mean	-4.222	-4.222
ml$concavity_mean	1.398	1.398
ml$concave.points_mean	2.142	2.142
ml$symmetry_mean	0.103	0.103
ml$fractal_dimension_mean	0.033	0.033
ml$radius_se	0.435	0.435

ml$texture_se	-0.007	-0.007
ml$perimeter_se	-0.023	-0.023
ml$area_se	-0.001	-0.001
ml$smoothness_se	15.854	15.854
ml$compactness_se	0.065	0.065
ml$concavity_se	-3.565	-3.565
ml$concave.points_se	10.568	10.568
ml$symmetry_se	1.697	1.697
ml$fractal_dimension_se	-7.146	-7.146
ml$radius_worst	0.195	0.195
ml$texture_worst	0.007	0.007
ml$perimeter_worst	-0.002	-0.002
ml$area_worst	-0.001	-0.001
ml$smoothness_worst	0.543	0.543
ml$compactness_worst	0.067	0.067
ml$concavity_worst	0.381	0.381
ml$concave.points_worst	0.464	0.464
ml$symmetry_worst	0.557	0.557
ml$fractal_dimension_worst	4.303	4.303

Table 7: Maximum Likelihood Estimate of Finite Mixture of Regression Model(MIXLASSO)

	Comp.1	Comp.2
ml$radius_mean	0.000	0.000
ml$texture_mean	0.000	0.000
ml$perimeter_mean	0.000	0.000
ml$area_mean	-0.000	-0.000
ml$smoothness_mean	0.000	0.000
ml$compactness_mean	0.000	0.000
ml$concavity_mean	0.000	0.000
ml$concave.points_mean	0.000	0.000
ml$symmetry_mean	0.000	0.000
ml$fractal_dimension_mean	0.000	0.000
ml$radius_se	0.000	0.000
ml$texture_se	0.000	0.000
ml$perimeter_se	0.000	0.000
ml$area_se	-0.000	-0.000
ml$smoothness_se	0.000	0.000
ml$compactness_se	0.000	0.000
ml$concavity_se	0.000	0.000
ml$concave.points_se	0.000	0.000
ml$symmetry_se	0.000	0.000
ml$fractal_dimension_se	0.000	0.000
ml$radius_worst	0.000	0.000
ml$texture_worst	0.000	0.000
ml$perimeter_worst	0.006	0.006
ml$area_worst	0.000	0.000
ml$smoothness_worst	0.000	0.000
ml$compactness_worst	0.000	0.000
ml$concavity_worst	0.000	0.000
ml$concave.points_worst	0.000	0.000
ml$symmetry_worst	0.000	0.000
ml$fractal_dimension_worst	0.000	0.000

Table 8: Maximum Likelihood Estimate of Finite Mixture of Regression Model(MIXSCAD)

	Comp.1	Comp.2
ml$radius_mean	0.000	0.000
ml$texture_mean	0.000	0.000
ml$area_mean	0.000	0.000
ml$smoothness_mean	-0.000	-0.000
ml$compactness_mean	-2.694	-2.694
ml$concavity_mean	0.000	0.000
ml$concave.points_mean	7.502	7.502
ml$symmetry_mean	0.000	0.000
ml$fractal_dimension_mean	-0.000	-0.000
ml$radius_se	0.000	0.000
ml$texture_se	0.000	0.000
ml$perimeter_se	0.000	0.000
ml$area_se	-0.000	-0.000
ml$smoothness_se	5.594	5.594
ml$compactness_se	0.000	0.000
ml$concavity_se	-0.000	-0.000
ml$concave.points_se	10.724	10.724
ml$symmetry_se	0.000	0.000
ml$fractal_dimension_se	-43.846	-43.846
ml$radius_worst	0.000	0.000
ml$texture_worst	0.000	0.000
ml$perimeter_worst	0.000	0.000
ml$area_worst	0.000	0.000
ml$smoothness_worst	0.000	0.000
ml$compactness_worst	0.000	0.000
ml$concavity_worst	0.000	0.000
ml$concave.points_worst	0.000	0.000
ml$symmetry_worst	0.000	0.000
ml$fractal_dimension_worst	10.030	10.030

Table 9: Maximum Likelihood Estimate of Finite Mixture of Regression Model(MIXHARD)

	Comp.1	Comp.2
ml$radius_mean	0.000	0.000
ml$texture_mean	0.000	0.000
ml$perimeter_mean	0.000	0.000
ml$area_mean	0.000	0.000
ml$smoothness_mean	-0.000	-0.000
ml$compactness_mean	-2.694	-2.694
ml$concavity_mean	0.000	0.000
ml$concave.points_mean	7.502	7.502
ml$symmetry_mean	0.000	0.000
ml$fractal_dimension_mean	-0.000	-0.000
ml$radius_se	0.000	0.000
ml$texture_se	0.000	0.000
ml$perimeter_se	0.000	0.000
ml$area_se	-0.000	-0.000
ml$smoothness_se	5.594	5.594
ml$compactness_se	0.000	0.000
ml$concavity_se	-0.000	-0.000
ml$concave.points_se	10.724	10.724
ml$symmetry_se	0.000	0.000
ml$fractal_dimension_se	-43.846	-43.846

ml$radius_worst	0.000	0.000
ml$texture_worst	0.000	0.000
ml$perimeter_worst	0.000	0.000
ml$area_worst	0.000	0.000
ml$smoothness_worst	0.000	0.000
ml$compactness_worst	0.000	0.000
ml$concavity_worst	0.000	0.000
ml$concave.points_worst	0.000	0.000
ml$symmetry_worst	0.000	0.000
ml$fractal_dimension_worst	10.030	10.030

7.2 Real Dataset 2

The data set contains variables on 1025 individuals. The outcome variable is target. The dataset has 12 covariates.

7.2.1 Data Description

The predictor variables are Age, Sex,Chest pain type (4 values), Resting blood pressure (in mm Hg on admission to the hospital), Serum cholestoral in mg/dl, Fasting blood sugar > 120 mg/dl, Resting electrocardiographic results (values 0,1,2), Maximum heart rate achieved, Exercise induced angina, Oldpeak = ST depression induced by exercise relative to rest, The slope of the peak exercise ST segment, Number of major vessels (0-3) colored by flourosopy, Thal: 0 = normal; 1 = fixed defect; 2 = reversable defect, and the response variable is Target : 1 = Disease, 0 = No Disease.

7.2.2 Data Visualization

Here are some visualizations that will help us to get started to see real results.

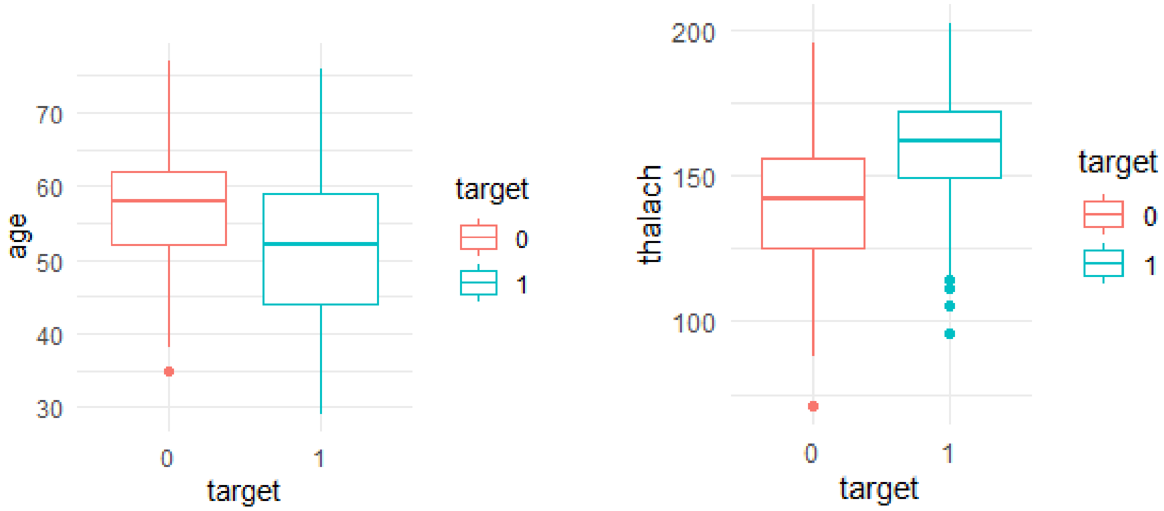

Figure 5: Boxplot of age and thalach vs target variable

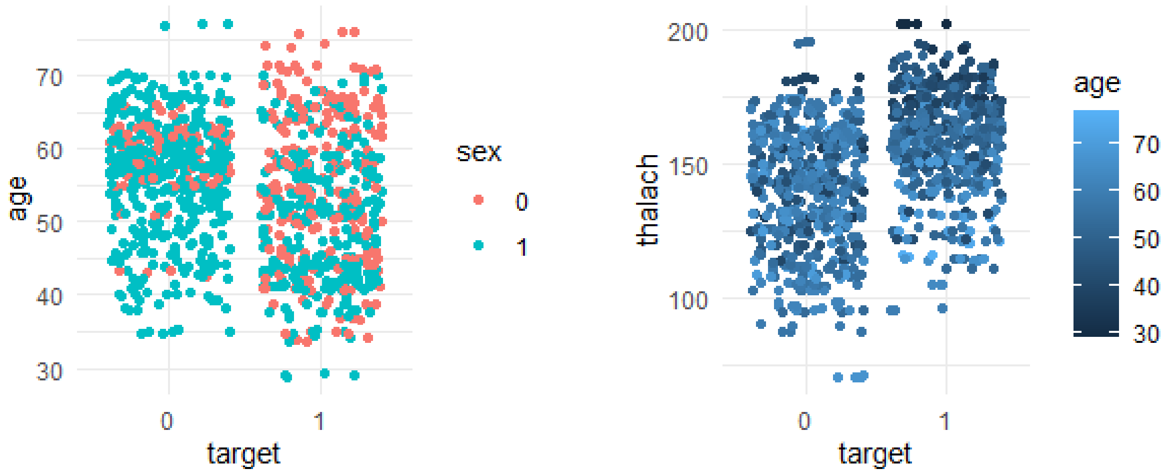

Figure 6: Plot age, sex vs target and thalach, age vs target

From the graphs we can say, among patients with a heart problem, men are mostly identified as women, especially when the age is going to be over 60 years.On average, the heart rate is higher in sick patients.

7.2.3 Model and Data Analysis

We fitted a multi-nomial logit FMR model addressed by Skrondal and Rabe-Hesketh (2004), with K = 2, corresponding to two segments, to the data arise from the program type conjoint analysis. Mathematically, the FMR model is given by

$$P(y_i) = P(Y_i = y_i) = (1-\pi)P_1(y_i) + \pi P_2(y_i)$$

Where

$$P_k(y_i) = \prod_{j=1}^{13}\prod_{a=1}^{2}[\frac{exp\{x_a^{\tau}\beta_k\}}{\sum_{l=1}^{3} exp\{x_l^{\tau}\beta_k\}}], k = 1,2; a = 1,2$$

The covariate x_a^{τ} is an 13×1 vector of dummy variables, corresponding to the ten attributes. Since the value of covariates x_a^{τ}'s did not change with subjects often enough, to make the parameters identifiable, an intercept term in the linear predictor can be excluded but we included it here.

First, We can consider subset selection method. The principle is simple: we calculate all the possible subspaces of predictors in order to see which one is the most efficient, according to your information criterion. This method uses the Leaps and bounds procedure Furnival and Wilson (1974) algorithm. The principle of Subset selection works relatively well when you have a predictor number, p, which does not exceed 30 or 40 because, since it is a "greedy algorithm", other methods will be more effective in terms of computational power required only this one. Here, our variable of interest "Target" is a binary variable and therefore requires classification, we cannot apply it properly. Same as backward and forward selection method is not feasible here.

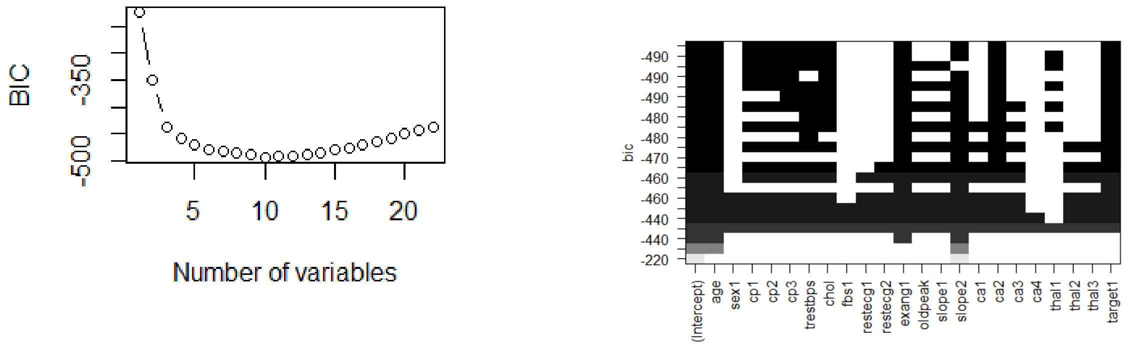

Figure 7: Subset selection method

Now, we are using shrinkage method, This is what we call regularization. The goal is very simple: we will penalize our estimators so that the coefficients will be close (Ridge) to zero or equal (Lasso). We can reduce the complexity of the model and have more efficient estimators. It is also a great way to select the most important predictors.

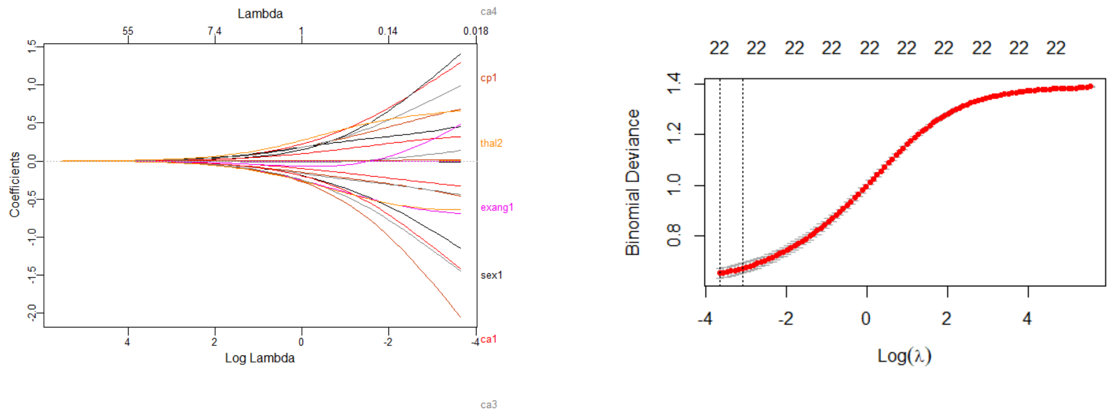

Figure 8: Ridge Estimation

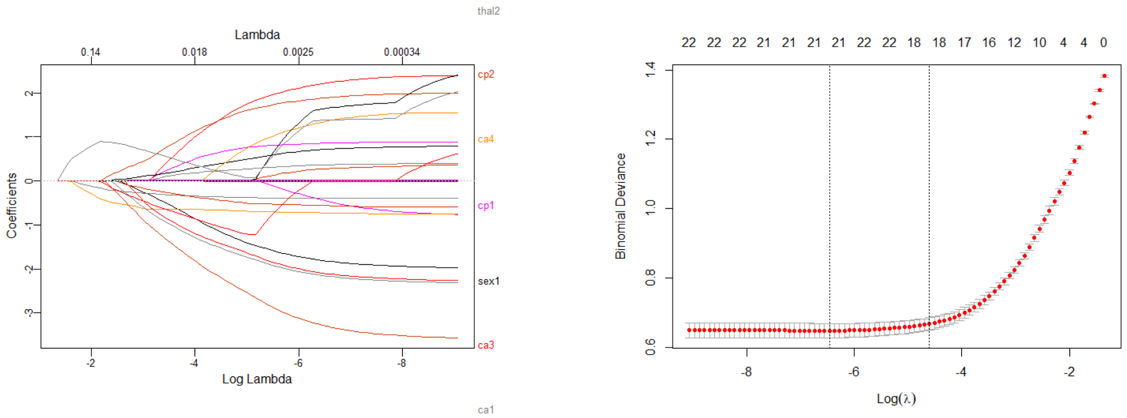

Figure 9: Lasso process

Table 10: Maximum Likelihood Estimate of Finite Mixture of Regression Model(MIXLASSO)

	Comp.1	Comp.2
age	-0.000	-0.000
sex1	0.000	-0.000
cp1	-0.000	-0.000
cp2	0.000	-0.000
cp3	0.000	0.000
trestbps	-0.000	-0.000
chol	-0.030	-0.011
fbs1	0.000	-0.000
restecg1	-0.000	-0.000
restecg2	-0.000	-0.000
thalach	0.000	0.000
exang1	0.000	0.000
oldpeak	0.000	-0.000
slope1	0.000	-0.000
slope2	0.000	-0.000
ca1	-0.000	0.000
ca2	0.000	0.000
ca3	0.000	-0.000
ca4	0.000	-0.000
thal1	0.000	0.000
thal2	0.000	0.000
thal3	0.000	-0.000

We compare with the graph of the ridge method, the difference is quite significant. The importance of the regressors does not necessarily change but the value of the coefficients is very different and will especially be much more penalized. Here, it would rather be necessary to select the variables Thal, CA and CP which remains a result relatively similar to the ridge method and our optimum $\lambda = 0.02596815$ for ridge and $\lambda = 0.001556749$ for lasso. Using mixture model proposed by Khalili and Chen (2007) in same data set we are finding that only chol has significant effect on target.

From those two real data analysis we can compared MIXLASSO, MIXSCAD and MIXHARD with new method proposed by Khalili and Chen (2007). From those we can see that MIXHARD and MIXSCAD chose model with more zero coefficients than the MIXLASSO penalty. In some cases we can see that one component has significant effect but another one has no effect over target variable.

8 Conclusion

Our reported paper introduced the penalized likelihood approach for variable selection in the context of finite mixture of regression. The penalty function is designed to be dependent on the size of the regression coefficients and the mixture structure. The new procedure is shown to be consistent in selecting the most parsimonious FMR model. It also reported a data adaptive method for selecting the tuning parameters and demonstrate the usage by extensive simulations. The new method with the MIXHARD and MIXSCAD penalty functions performed as well as the BIC method while it is computationally much more efficient. In addition, as in the example of market segmentation application, the new method can also be used to suggest a set of plausible models to be examined by the BIC method if desired. This helps to reduce the computational burden of using the BIC, substantially.

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.

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.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, 16(4):499–511.

Hennig, C. (2000). Identifiablity of models for clusterwise linear regression. *Journal of Classification*, 17(2).

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.

McLachlan, G. J. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

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. (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.