Model Selection in Multivariate Analysis with Missing Data

BY

FEI SHI
B.S., Sun Yat-Sen University, China, 2004
M.S., University of Illinois at Urbana-Champaign, Urbana, 2005

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Public Health Sciences
in the Graduate College of the
University of Illinois at Chicago, 2012

Chicago, Illinois

Defense Committee:
Huayun Chen, Chair and Advisor
Donald Hedeker
Sally Freels
Hui Xie
Yang Dai, Bioengineering

This thesis is dedicated to my father Shi, Yaohua,

my mother Zhu, Lili.

ACKNOWLEDGMENTS

I am very grateful to my dissertation advisor Professor Huayun Chen, for his inspiring guidance and enthusiastic support from development to completion of this dissertation. Without his unique insights, broad knowledge and constant encouragement in my research, I could not have achieved anything that I have done here. The past two years that we worked together are valued memories of mine.

I also want to thank Professor Donald Hedeker, Sally Freels, Yang Dai and Hui Xie, who have served on my dissertation committee. Their helpful comments and suggestions contributes to the better quality of the dissertation and strengthen my understanding in statistics.

I want to thank my parents, who give me constant support and encouragement. Their patience and confidence on me motivates me to accomplish my research goals. Last but not least, I want to thank my girl friend Hoiyan Mui, who is always there to support me through happiness and frustration in each day of my graduate study.

FS

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

TABLE **PAGE**

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion.
BIC	Bayesian Information Criterion.
BMI	Body Mass Index.
EM	Expectation-Maximization.
FCS	Fully Conditional Specification.
GCV	Generalized Cross-Validation.
GQ	Gaussian Quadratures.
HT	Height.
ITS	Impute Then Select.
LARS	Least Angle Regression.
LASSO	Least Absolute Shrinkage and Selection Operator.
LLA	Local Linear Approximation.
LSA	Least Square Approximations.
MAR	Missing At Random.
MCAR	Missing Completely At Random.
MCMC	Markov Chain Monte Carlo.
MICE	Multivariate Imputation By Chained Equations.
ML	Maximum Likelihood.
MLE	Maximum Likelihood Estimator.

LIST OF ABBREVIATIONS (Continued)

MPLE	Maximum Penalized Likelihood Estimates.
MRME	Median of Ratio of Model Error.
NI	Non-Ignorable.
NN-Garotte	Non-Negative Garotte.
OLS	Ordinary Least Squares.
RSS	Residual Sum of Squares.
SCAD	Smoothly Clipped Absolute Deviation.
S.D	Standard Deviation.

SUMMARY

In this dissertation, a new model selection algorithm based on maximizing penalized likelihood function with the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) penalty function was developed for missing data problems. Current model selection method in missing data problems iteratively optimizes the penalized Q-function in Expectation Maximization (EM) algorithm which is a computationally expensive process. We proposed a new model selection algorithm that utilized an approximation based on information identity to the observed data log-likelihood to obtain the maximum penalized likelihood estimate (MPLE). A modified tuning parameter criterion based on BIC (Schwarz, 1978) for missing data problems was proposed to select the optimal tuning parameter for the penalty function. Furthermore, we proposed a new model selection scheme that not only selects covariates for the outcome variable but also selects covariate models, which are important in high dimensional covariates subject to missing values.

Following Fan and Li (2001), we proved the existence and consistency of the proposed maximum penalized likelihood estimators. The current method for selecting the optimal tuning parameter is based on Q-function in place of the observed data log-likelihood, which may cause an over-fit effect. By using Taylor expansion, we can rewrite the observed data log-likelihood asymptotically in a least square form so that a good approximation to the BIC criteria can be obtained and the efficient path finding algorithm least angle regression (LARS) can be applied to find the sparse MPLE, using local linear approximation proposed in Zou and Li (2008). We have implemented the proposed model selection algorithm in linear regression and logistic regression models. Monte Carlo simulation with rejection sampling was used to approximate the intractable integrals in computing the expected full-data log-likelihood function conditional on

the observed data (Q-function) in EM algorithm for logistic regression models. Three computer programs are developed for model selection in logistic regression with missing continuous, binary and mixed continuous and binary covariates data, respectively.

Several simulations were carried out to examine the performance of the proposed algorithm. Results show that the proposed method can dramatically reduce the model error and consistently identify the true model with large sample sizes. Our proposed algorithm outperforms other model selection methods for missing data problems in the simulation studies in identifying a larger proportion of correct-fits as sample size increases.

Data from a case-control study to investigate the potential risk factors of hip fracture among male veterans were used to illustrate the application of the proposed method in model selection. We ran several selection processes on the data with the proposed and imputation methods. Results show that only 4 out of 27 covariates are selected as significant risk factors to predict the presence of hip fracture, while 15 are selected by traditional step-wise selection on a complete-case analysis.

CHAPTER 1

A MOTIVATING EXAMPLE WITH INCOMPLETE DATA

The hip fracture data were from a case-control study conducted at the University of Illinois at Chicago (Barengolts et al., 2001), to investigate potential risk factors of hip fracture among male veterans. There are 218 cases and 218 controls with each case matched with a control on age and race. In total, there are 27 risk factors recorded. Among the 27 covariates, 10 of them are continuous. The rest are binary. Arithmetic mean and standard deviation (S.D) stratified by case-control status for continuous variables are listed in Table I. From Table I, we can see mean and s.d for the matching variables are very similar, suggesting the matching was well executed. Table II lists the frequencies of the binary variables. We can see from Table II that, for most variables, cases are more likely subject to missing values than controls.

To analyze the data, a logistic regression model with hip fracture status as the binary outcome and potential risk factors as predictors is used. To see which of the 27 potential risk factors contribute to the risk of hip fracture, model selection is essential. Traditionally, variable selection methods, such as forward, backward, and stepwise selections are usually used to select a subset of covariates by some selection criteria, such as residual sum of squares (RSS), Akaike information criterion (AIC) or Bayesian information criterion (BIC). One disadvantage of the traditional model selection methods (Breiman, 1996) is that they separate selection and estimation processes. As a result, the estimator suffers from uncertainty or instability. Many new variable selection approaches that perform selection and estimation simultaneously have since been proposed. Among them, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) in linear regression to minimize the RSS subject to the constraint that the sum of the absolute value of the coefficients is less than a constant. The constraint

TABLE I

ARITHMETIC MEAN OF ALL CONTINUOUS VARIABLES

Variable	Group	Observed	Missing	Mean	$s.d$	P-value
Age	Case	218	0	69.9	11.2	
	Control	218	0	69.6	10.5	0.76
Weight	Case	190	28	68.8	14.5	
	Control	204	14	82.3	18.7	< 0.0001
Height	Case	191	27	148.5	62.9	
	Control	204	14	149.1	60.7	0.95
BMI	Case	189	29	22.6	4.7	
	Control	203	15	27.2	6.1	< 0.0001
Albumin	Case	170	48	3.4	0.7	
	Control	163	55	3.8	0.6	< 0.0001
Cholesterol	Case	170	48	167.6	47.5	
	Control	181	37	193.5	40.7	< 0.0001
hgb	Case	190	28	12.0	2.1	
	Control	193	25	13.6	1.8	< 0.0001
hct	Case	191	27	35.7	6.5	
	Control	203	15	40.2	5.5	< 0.0001
BUN	Case	190	28	23.3	17.7	
	Control	205	13	18.0	8.3	0.0002
Cr	Case	190	28	1.8	2.9	
	Control	206	12	1.4	1.5	0.09

TABLE II

ARITHMETIC MEAN OF ALL BINARY VARIABLES

Variable	Group	Observed	Missing	Exposed(Unexposed)	P-value
Race	Case	218	0	137 (81)	
	Control	218	0	137 (81)	1
Etoh	Case	179	39	109 (70)	
	Control	213	5	67 (146)	< 0.0001
Smoke	Case	172	46	118 (54)	
	Control	210	8	79 (131)	< 0.0001
CVA	Case	205	13	40 (165)	
	Control	215	3	26 (189)	0.04
Dementia	Case	204	14	45 (159)	
	Control	218	0	10 (208)	< 0.001
Parkinson	Case	204	14	11 (193)	
	Control	218	0	4 (214)	0.05
Seizure	Case	200	18	25 (175)	
	Control	217	1	6 (211)	0.0002
Sedat	Case	198	20	30 (168)	
	Control	216	2	23 (193)	0.17
NSAIDS	Case	198	20	56 (142)	
	Control	216	2	113 (103)	< 0.0001
Steroids	Case	195	23	6 (189)	
	Control	210	8	3 (207)	0.26
Lasix	Case	197	21	32 (165)	
	Control	216	2	47 (169)	0.15
HCTZ	Case	197	21	20 (177)	
	Control	216	2	45 (171)	0.003
Antiseiz	Case	197	21	33 (164)	
	Control	216	2	5 (211)	< 0.0001
CaCO3	Case	197	21	16 (181)	
	Control	216	2	10 (206)	0.14
LevoT4	Case	186	32	11 (175)	
	Control	210	8	11 (199)	0.77
AntiChol	Case	186	32	10 (176)	
	Control	203	15	41 (162)	< 0.0001
COPD	Case	201	17	42 (159)	
	Control	211	7	28 (183)	0.04

TABLE III

MISSING PATTER FOR SELECTED VARIABLES (PARTIAL)

Total	BMI	Hgb	Albumin	Etoh	Smoke	Dementia	Antiseiz	...	Cholesterol
232	1	1	1	1	1	1	1	...	1
17	0	1	1	1	1	1	1	...	1
12	1	0	1	1	1	1	1	...	1
21	1	1	0	1	1	1	1	...	1
1	0	1	0	1	1	1	1	...	1
7	1	1	1	1	0	1	1	...	1
1	0	1	1	1	0	1	1	...	1
17	1	1	1	0	0	1	1	...	1
2	0	1	1	0	0	1	1	...	1
1	1	0	1	0	0	1	1	...	1
1	1	1	0	0	0	1	1	...	1
5	1	1	1	1	1	1	1	...	1
1	0	1	1	1	1	1	1	...	1
1	1	1	0	1	1	1	1	...	1
15	1	1	1	1	1	1	1	...	1
1	1	1	1	0	0	1	1	...	1
2	1	0	1	1	1	1	1	...	1
1	1	1	1	1	0	1	1	...	1
1	1	1	1	1	1	0	0	...	1
1	1	1	1	1	0	1	0	...	1
1	0	1	1	1	0	1	0	...	1
1	1	1	1	0	0	1	0	...	1
1	1	1	1	0	0	0	0	...	1
1	1	1	1	1	1	1	1	...	1
4	1	0	1	1	1	1	1	...	1
1	1	1	0	1	1	1	1	...	1
1	1	1	0	1	1	1	1	...	1
1	1	0	1	1	1	1	0	...	1
					...				
1	0	0	0	0	0	1	1	...	0
1	1	0	0	1	1	1	1	...	0
1	1	0	0	1	1	1	0	...	0
1	0	0	0	0	1	1	0	...	0
3	0	0	0	0	0	1	0	...	0
1	1	0	0	0	0	0	0	...	0
1	1	0	0	1	1	0	0	...	0
1	1	0	0	0	1	0	0	...	0
7	1	0	0	0	0	0	0	...	0
2	0	0	0	0	0	0	0	...	0

defined by L_1 Euclidean norm leads to the exact zero estimation for some coefficients so that it selects variables and estimate regression coefficients simultaneously. Fan and Li (2001) studied LASSO carefully and concluded that L_1 penalty is the only convex penalty function that can generate a continuous sparse solution within the L_p function family. It further proposed a new penalty function called SCAD (Smoothly Clipped Absolute Deviation Penalty) and showed it can select the correct sparse model with a suitable choice of the tuning parameter for the penalty function as sample size increases. To apply these variable selection approaches to the hip fracture data, one major challenge is the presence of missing values in many of the potential risk factors. All 27 variables, except the matching variables age and race, are subject to missing values. Only 227 out of 436 subjects (52.1%) have complete records for all the 27 covariates. Altogether, there are 74 missing-data patterns, 63 of them have fewer than 5 observations. Table III summarizes the missing data patterns. In the presence of missing covariates, conventional model selection approaches can only apply to the complete cases, which may yield biased estimator and significantly reduce the power. Most newly developed variable selection approaches in the literature cannot satisfactorily solve the problem of variable selection with missing data.

We did a preliminary analysis of the data using the step-wise selection for completely observed cases. The selection picks a set of 15 covariates as potential risk factors for hip fracture: body mass index (BMI), hgb, albumin, etoh, smoke, dementia, Antiseiz, LevoT4, AntiChol, CVA, Lasix, HCTZ, BUN, cholesterol and height (HT).

In this dissertation, we will extend the penalized likelihood approach to generalized linear models with missing covariates, particularly to multiple linear and logistic regression models. We will extend model selection to all covariates models that are specified in missing covariate problems. Several issues need to be addressed. The first is how to maximize the

observed data log-likelihood along with SCAD penalty function to select important variables and to estimate parameters since the observed data log-likelihood often involves intractable integration and is not in a closed form. The second is to select the appropriate tuning parameters to produce a consistent sparse coefficient estimator. Ibrahim et al. (2008), proposed to optimize a selection criterion, called IC_Q statistics, which is an approximation to the observed data log-likelihood by the expectation of complete data log-likelihood conditional on observed data at the maximum likelihood estimator (MLE): the Q-function in the EM algorithm. It has been showed in Ibrahim et al. (2008), that a model selection criterion based on the Q-function alone can overstate the amount of information in the missing data compared with the observed data log-likelihood function. Thus, in modeling multivariate linear regression on data with missing covariates, we will directly use observed data log-likelihood in the BIC, which has been showed in Wang et al. (2007), that it can consistently identify the true model. For logistic regression, we will replace IC_Q criterion by a better approximation to the observed data information matrix in BIC.

CHAPTER 2

VARIABLE SELECTION FOR FULLY OBSERVED DATA

2.1 Regression Models and Their Estimation

Let Y denote the outcome variable of interest and $(X_1, \cdots, X_p)$ be a p-dimension covariate vector. To answer research questions such as whether $(X_1, \cdots, X_p)$ is useful for predicting Y or whether a subset of of $(X_1, \cdots, X_p)$ is associated with Y, a commonly used regression model has the form:

$$Y = m(X_1, \cdots, X_p) + \varepsilon,$$

where $m(X_1, \cdots, X_p)$ is the regression function, usually the conditional expectation of Y given $(X_1, \cdots, X_p)$, $E(Y|X_1, \cdots, X_p)$, and ε is a random error with mean 0 and finite variance. In linear regression,

$$m(X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Other examples of regression models include the generalized linear model (Nelder, 1972) such as the logistic regression and Poisson regression. These models are flexible generalizations of the linear regression models and can be applied to different kinds of data, such as the logistic regression models for binary responses, Poisson models for counts.

Suppose that the observed data $(X_{i1}, \cdots, X_{ip}, Y_i), i = 1 \ldots n$ are independent identically distributed copies of $(X_1, \cdots, X_p, Y)$. Let $\mathbf{Y} = (Y_1, \cdots, Y_n)^t$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \cdots, \mathbf{X}_{ip})$, and

$\mathbf{X} = (\mathbf{X}_1^t, \cdots, \mathbf{X}_n^t)^t$. For linear regression models, the least squares estimate is obtained via minimizing

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\beta)^2 &= \|\mathbf{Y} - \mathbf{X}\beta\|^2 \\ &= \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 + \|\mathbf{X}(\hat{\beta} - \beta)\|^2,\end{aligned}$$

where $\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}$ is the ordinary least squares estimate and $\|\cdot\|$ is the L^2 Euclidean norm. The problem is equivalent to minimizing with respect to β,

$$(\hat{\beta} - \beta)^T(\mathbf{X'X})(\hat{\beta} - \beta).$$

For generalized linear models, if $f(Y;\theta)$ is the density function for the observation Y given the parameter θ, then the log-likelihood function, expressed as a function of the mean-value parameter $E(Y|X_1, \cdots, X_p) = \mu(\theta) = g^{-1}(\mathbf{x}\beta)$ is

$$l\{\mu(\beta); \mathbf{Y}\} = \sum_{i=1}^{n} \log f(Y_i; \theta_i), \tag{2.1}$$

where $g(\mu)$ is the link function. The maximum likelihood estimator for β may be obtained through maximizing $l\{\mu(\beta); \mathbf{Y}\}$ with respect to β.

In many biological and medical studies, a large number of covariates are collected. One of the important tasks is to determine which covariates are truly associated with the outcome and which variables are not. It has been demonstrated that including a very large number of covariates in the regression model not only significantly increase the difficulties in interpreting the data (Breiman, 1996), but also decrease the accuracy in prediction (Roecker, 1991). Hence,

variable selection is very important in regression analysis for better interpreting the data and for achieving a smaller prediction error, especially, when a large number of covariates are involved.

2.2 Traditional Variable Selection Approaches

The traditional variable selection approaches for linear and nonlinear regression models select a subset of covariates that fit the data well in the sense of minimizing measures of goodness-of-fit. Those measurements include but are not limited to the least-squares, the coefficient of determination R^2, the AIC or a log-likelihood. The most commonly used procedures of variable selection include forward selection, backward elimination, stepwise selection and best subset selection. In the following, we give a description for these selection procedures in linear models. Similar logic can be applied to nonlinear models or with other selection criteria. Forward selection begins by selecting a single predictor variable that produces the best fit with regard to an inclusion criterion, e.g., the smallest residual sum of squares. The process of forward selection can be implemented in details as follows. For a given variable X_j, we minimize

$$S(b_j) = \sum_{i=1}^{n} (Y_i - b_j X_{ij})^2$$

with respect to b_j. It is easy to see that the minimizer is given by

$$\hat{b}_j = \sum_{i=1}^{n} X_{ij} Y_j / \sum_{i=1}^{n} X_{ij}^2.$$

Subsititute the estimator back into previous equation, we have

$$S(\hat{b}) = \sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} X_{ij} Y_j)^2 / \sum_{i=1}^{n} X_{ij}^2.$$

The first variable we select is the one which minimizes the second term of above equation:

$$(\sum_{i=1}^{n} X_{ij}Y_j)^2 / \sum_{i=1}^{n} X_{ij}^2.$$

Let the variable selected be X_{j_1}. We then perform simple linear regression of y on X_{j_1}. Next consider the residual vector as the response and project the other predictors orthogonal to X_{j_1} and repeat the selection process. After k steps, a set of predictors X_{j_1},X_{j_2},...,X_{j_k} are selected. A stop criterion can be set up. For example, in some statistical programs, a F test is performed in each step. When the calculated P-value is smaller than the pre-specified cut-point the selection is stopped. Breiman (1996) showed that the search may not find the best model. In addition, a drawback of the forward selection is its instability: a relatively small change in the data might cause one variable to be selected instead of another, after which subsequent choices may be completely different.

Similarly, backward elimination starts with all p variables in the model and sequentially removes variables that contribute least to the fit. Let RSS_p be the corresponding residual sum of squares for the model have all variables. The variable that yields the smallest value of RSS_{p-1} is deleted from the model. The process continues until there is no variable left or, until the stopping criterion is satisfied.

Stepwise selection (Efroymson, 1960) is a variation of the forward selection. Its first step is the same as the forward selection. After the first step, it adds the covariate that has the smallest RSS with corresponding F-test significant to the specified cut-off value. Then backward selection is performed until no variable can be removed based on the exclusion criterion. Hence, it incorporates criteria for addition and deletion of variables. Implement details can be seen in Miller (1990). Stepwise selection has the advantage of computation convenience and easy

to interpret since it results in a sparse model. This method may not find the best model. In addition, one major drawbacks is its instability, i.e., "A small change in data would result in a very different model being selected so that lower the prediction accuracy" (Tibshirani, 1996, p267).

Subset selection performs an exhaustive search over all possible subsets of covariates. One or a small number of the more promising models may be selected on some criteria. This approach is better in that it performs an exhaustive search to locate the best model. However, the exhaustive search can be costly. The number of possible subsets of one or more variables out of p is $2^p - 1$. Thus, the computational cost roughly doubles with each additional variable. Besides it is expensive in computation, it also suffers from extremely instability, i.e., a small change in the data would largely change the variable selection result, according to Breiman (1996).

In summary, classical variable selection that are computationally feasible may not yield consistent model. For those selections yielding consistent selection, such as the subset selection, computation can be prohibitive.

Much progress have been made in the past decades to variable selection techniques. First, a more general penalized least squares regression, bridge regression, was introduced in Frank and Friedman (1993), by using penalty functions $\sum |\beta_j|^\gamma$, when $\gamma > 0$. This is an improvement from the ridge regression previously proposed by Hoerl and Kennard (1970a, 1970b), which sets $\gamma = 2$. Tibshirani (1996) proposed the LASSO with $\gamma = 1$ or the L_1 norm. The LASSO penalty enjoys the properties of continuity and sparsity, i.e., its solutions is continuous with respect to tuning parameter value and it can estimate the regression coefficients by exact zeros.

2.3 Variable Selection via Least Absolute Shrinkage and Selection Operator

2.3.1 Ridge Regression

To overcome the ad-hoc and instability nature of classical variable selection methods, ridge regression was suggested as a strong competitor to subset regression in terms of variance reduction (Hoerl and Kennard 1970a, 1970b). It minimizes a penalized sum of squares

$$\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where λ is a shrinkage parameter. It reduces to the ordinary least-squares regression when $\lambda = 0$. Increasing λ shrinks the coefficient estimates, but none are set equal to zero. Since it does not set any coefficients to 0, its solution has the same complexity as the ordinary least square (OLS) estimator and may not be easily interpretable. In contrast, LASSO that uses of an L_1 penalty instead of an L_2 penalty can yield exact zero regression coefficient estimates.

2.3.2 Variable Selection in Linear Regression Models

To solve aforementioned problems, alternative methods were sought. Breiman(1995) proposed non-negative garotte (NN-Garotte) method which minimizes

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} c_j \tilde{\beta}_j^0 X_{ij} \right)^2$$

with constraints $c_j \geq 0$ and $\sum_{j=1}^{p} c_j \leq s$, where $\tilde{\beta}_j^0$ are ordinary least square estimates. The nn-garotte selection is a more stable procedure which in general shrinks more coefficients to zeros than classical subset selection. A drawback of nn-garotte method is its use of the OLS estimator in its objective function because the garotte estimates may suffer when the OLS estimator performs poorly, which usually happens with data having collinearity problems. When the

number of predictor variables is comparable to the sample size, the OLS estimators may not be even unique.

Motivated by Breiman's NN-Garotte method, Tibshirani (1996) proposed the LASSO. It minimizes the sum of squared errors subject to a constraint on the L_1 norm of the regression coefficients. The LASSO estimates $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ is defined by

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 \right\} \tag{2.2}$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$. The LASSO approach is a special case of the more general "bridge regression" approach proposed by Frank and Friedman (1993) which minimizes the squared errors subject to the constraint $\sum_j |\beta_j|^q \leq t$ for a given $q \geq 0$. The LASSO penalty function corresponds to $q = 1$. Two other prominent special cases are ridge regression where $q = 2$ and the best subset selection where $q = 0$.

Consider the special case where $x_{i1}, \ldots, x_{ip}, i = 1, \ldots, n$ are standardized and orthogonal. That is, $\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} x_{ij}x_{ik} = 0, j \neq k$ and $\sum_{i=1}^{n} x_{ij}^2 = 1$. Assume also that $\sum_{j=1}^{n} y_j = 0$, then

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 &= \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 + \\ &\quad \sum_{i=1}^{n} \left\{ (\hat{\beta}_1 - \beta_1) x_{i1} + \ldots + (\hat{\beta}_p - \beta_p) x_{ip} \right\}^2, \end{aligned} \tag{2.3}$$

where $\hat{\beta}_j = (\sum_{i=1}^{n} x_{ij}^2)^{-1} \sum_{i=1}^{n} x_{ij} y_i$. Thus, minimizing a least square with a LASSO type penalty as follows

$$\sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

is equivalent to minimizing

$$\sum_{i=1}^{n}\Big\{(\hat{\beta}_1-\beta_1)x_{i1}+\ldots+(\hat{\beta}_p-\beta_p)x_{ip}\Big\}+\lambda\sum_{j=1}^{p}|\beta_j|. \tag{2.4}$$

Note that

$$\sum_{i=1}^{n}\Big\{(\hat{\beta}_1-\beta_1)x_{i1}+\ldots+(\hat{\beta}_p-\beta_p)x_{ip}\Big\}=\sum_{j=1}^{p}(\beta_j-\hat{\beta}_j)^2.$$

Therefore the original LASSO problem becomes

$$\arg\min_{\beta}\left\{\sum_{j=1}^{p}(\beta_j-\hat{\beta}_j)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{2.5}$$

which is equivalent to minimizing $\Big\{(\beta_j-\hat{\beta}_j)^2+\lambda|\beta_j|\Big\}$ for each $j=1,\ldots,p$.

Let us consider the penalized least square problem given by

$$\frac{1}{2}(z-\theta)^2+\lambda|\theta|. \tag{2.6}$$

Its solution is given by the threshold rule as

$$\hat{\theta}=\mathrm{sgn}(z)(|z|-\lambda)_+ \tag{2.7}$$

from Donoho and Johnstone (1994). Figure 1 plots the above least square problem with different z values when $\lambda=2$. We can see that L_1 penalty function shrinks the minimum value of θ obtained at which θ is close to zero. Plotting the thresholding rule in Equation 2.7 gives us a visual insight that LASSO automatically sets small estimated coefficients in the least square estimator to zeros. Further, Fan and Li (2000) gave a sufficient condition for resulting estimator

to be a sparse one, that is "the minimum of the function $|\theta| + p'_\lambda(|\theta|)$ is positive." We can easily verify that LASSO satisfies this requirement but not for the other bridge regression with $q > 1$.

Fan and Li (2000) pointed out that the solution of bridge regression is continuous only when $q \geq 1$. On the other hand, only when $q \leq 1$ can yield exact zero regression coefficient. When $q > 1$, shrinkage does not lead zero regression coefficient. Hence, the only continuous solution with a thresholding rule is the L_1 penalty.

To find the bridge regression coefficients, Fu (1998) designed a modified Newton-Raphson algorithm for the case $q > 1$ and the shooting algorithm for the LASSO estimates. The shooting algorithm is very attractive in computation speed and memory since it has a convergence rate of $p \log p$, in contrast with the convergence speed of 2^p for the quadratic programming method proposed by Tibshirani (1996).

2.3.3 Least Angle Regression

Efron et al. (2004), proposed a new model selection algorithm: LARS, which can be used to get LASSO solutions. It is highly computationally efficient because it requires the same order of magnitude of computational effort as the ordinary least squares applied to the full set of covariates (Efron et al., 2004). It provides a convenient way to efficiently calculate adaptive LASSO and an approximation algorithm for non-concave penalized likelihood selection methods.

The idea of LARS algorithm is as follows. In the first step, the variable with the largest correlation with the outcome is selected. However, the regression coefficient of the first selected variable is not set as high as in the traditional regression model so that the residual is no longer correlated with the first selected variable. Instead, the regression coefficient for the first selected variable is chosen such that the residual correlation with the first selected variable is reduced to the level of the maximum of the correlations of the residual with all other unselected

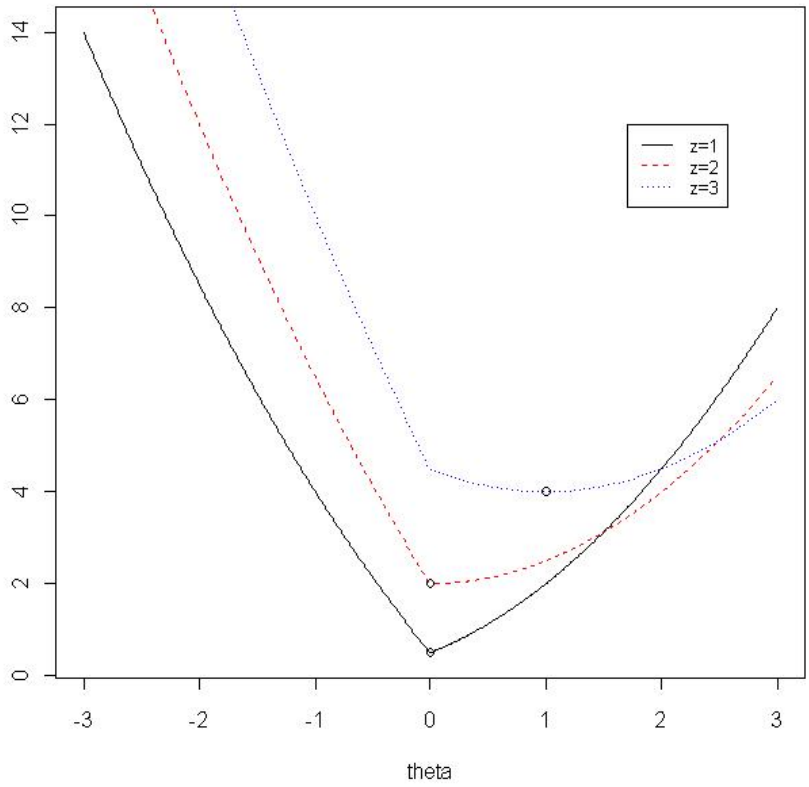

Figure 1. Plot of least square function with LASSO penalty function when $\lambda = 2$

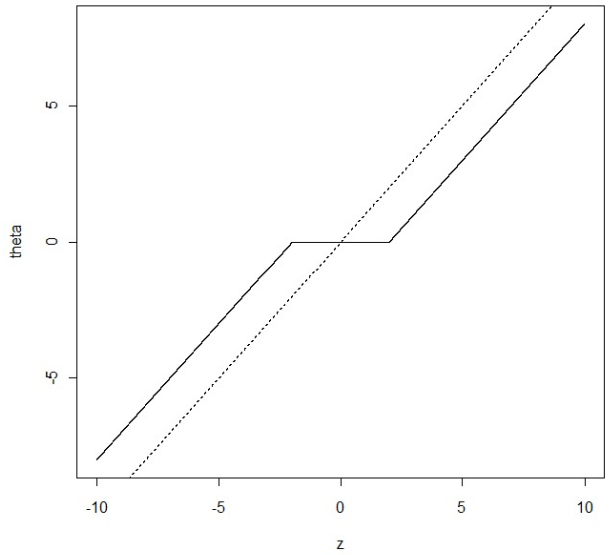

Figure 2. Plot of LASSO thresholding function with $\lambda = 2$

variables. This determines both the regression coefficient for the first selected variable and the next variable to be entered. After that, LARS will go in a direction "equiangular" with the current active set until a third variable has the same correlation with current residual as the current active set. The process continues until all covariates enter into the active set.

Least angle regression algorithm does not directly give LASSO solutions since LASSO solutions has a restriction that any non-zero coefficient must have the same sign as the corresponding current correlation with the residual. But a minor modification of LARS algorithm can produce LASSO estimates. As described in Efron et al. (2004), the detailed modified LARS algorithm for LASSO solutions can be summarized as follows: Suppose covariate

vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ are linearly independent. For A a subset of indices $\{1, 2, \ldots, p\}$, define the matrix

$$X_A = (\ldots s_j \mathbf{x_j} \ldots)_{j \in A}$$

where the signs s_j equal ± 1. Let $g_A = X_A' X_A$ and $A_A = (\mathbf{1_A'} g_A^{-1} \mathbf{1_A})^{-1/2}$, where $\mathbf{1_A}$ is a vector of 1's of length equaling $|A|$, the size of A. Let $\mathbf{u_A} = X_A w_A$ where $w_A = A_A g_A^{-1} \mathbf{1_A}$.

Algorithm of LASSO

1. Let $\mu_\mathbf{A}$ is the current LARS estimate. Begin with $\hat{\mu}_\mathbf{0} = \mathbf{0}$. Let $\hat{\mathbf{c}} = X'(\mathbf{y} - \hat{\mu}_\mathbf{A})$ and $\hat{C} = \max_j\{|\hat{c}_j|\}$. Add j into A, where $j = \arg\min\{|\hat{c}_j| = \hat{C}\}$.

2. Letting $s_j = sign\{\hat{c}_j\}$, we update X_A, A_A and $\mathbf{u_A}$ and calculate $\mathbf{a} = X'\mathbf{u_A}$.

3. Calculate

$$\hat{\gamma} = \min_{j \in A^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{A_A - a_j}, \frac{\hat{C} + \hat{c}_j}{A_A + a_j} \right\}$$

   where $\hat{j}$ is the minimizing index and min^+ means the minimum is taken over only positive components within each choice of j. Calculate $\hat{\beta}_j = \hat{\beta}_j^{k-1} + \hat{\gamma}\hat{d}_j$ for $j \in A$.

4. Let

$$\hat{\mathbf{d}} = \begin{cases} s_j(w_A)_j & \text{when } j \in A \\ 0 & \text{elsewhere.} \end{cases}$$

   If $\tilde{\gamma} \equiv \min^+\{\frac{\hat{\beta}_j}{\hat{d}_j}\} < \hat{\gamma}$, update $A_+ = A - \tilde{j}$, $\hat{\beta}_j = \hat{\beta}_j^{k-1} - \tilde{\gamma}\hat{d}_j$ for $j \in A$ and $\hat{\mu}_{\mathbf{A}+} = \mathbf{x}\hat{\beta}$ where $\tilde{j} = \arg\min^+\{\frac{\beta_j}{\hat{d}_j}\}$ then goes to step 1; otherwise continue to step 5.

5. Updates $\hat{\mu}_{\mathbf{A}}$ to

$$\hat{\mu}_{\mathbf{A}+} = \hat{\mu}_{\mathbf{A}} + \hat{\gamma}\mathbf{u}_{\mathbf{A}}$$

and $A_+ = A \cup \{\hat{j}\}$. Go to step 1.

Repeating this iteration procedure will produce a total r sets of coefficient estimates, where r represents total steps of thresholding value of t in LASSO when there is new covariate enter into regression or remove from it. One of the most remarkable contributions of LARS is its speed. From Efron et al. (2004, P.443), "The entire sequence of LARS steps with $p < n$ variables requires $O(p^3+np^2)$ computations the cost of a least squares fit on p variables." Least angle regression provides an efficient and simple way to do variable selection through LASSO.

2.3.4 Selection of Tuning Parameter

The modified LARS algorithm can be used to obtain the complete solution path for LASSO in the previous section. As a result, a LASSO solution can be determined once the bound t is given. In practice, however, one often needs to determine the tuning parameter t based on the data. Here we describe three commonly used methods to estimate the LASSO tuning parameter t for linear regression models: five-fold cross-validation, generalized cross-validation and BIC-type selection criterion (Schwarz,1978). In LASSO, the optimization problem (Equation 2.2) can be rewritten as a Lagrangian problem below

$$\hat{\beta}_p = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2.8}$$

where $\lambda \geq 0$ and has an one-to-one relationship with L_1 constraint parameter t. From Osborne et al. (2000), for a given LASSO solution $\hat{\beta}$ that minimizes Equation 2.8, λ can be calculated as $\lambda = \|\mathbf{X^T}(\mathbf{y} - \mathbf{X}\hat{\beta})\|_\infty$, where $\|\cdot\|_\infty$ is defined the maximum L_p norm as $\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$.

Let θ be the tuning parameter to be estimated. In LASSO problems, $\theta = \lambda$ or t. Fivefold cross-validation procedure for linear regression is as follows. Split the full dataset into 5 parts. For kth part ($k = 1, \ldots, 5$), fit the model to the other 4 parts of data and calculate the prediction error of the fitted model using the kth part of the data. For linear regression, prediction error is usually defined as follow

$$E(\mathbf{Y} - \mathbf{X}\hat{\beta})^2 = (\hat{\beta} - \beta)^T E(\mathbf{X^T X})(\hat{\beta} - \beta) + \sigma^2 I,$$

where σ^2 is the residual variance. Find λ that minimizes the prediction error. The criterion for generalized cross-validation is the average of weighted residual sum of square for linear regression given by

$$GCV(\theta) = \frac{1}{n} \frac{\|\mathbf{Y} - \mathbf{X}\beta(\theta)\|^2}{\{1 - p(\theta)/n\}^2},$$

where $p(\theta)$ is the approximation of number of effective parameters in the constrained solution of LASSO given by

$$p(\theta) = tr\{\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\mathbf{X}^T\},$$

where $\mathbf{W} = diag(|\hat{\beta}_j|)$. In Fu (1998), the effective number of parameters of the model is computed as $(p(\theta) - n_0)$, where n_0 is the number of zero components in β's and $\mathbf{W} = diag(2|\hat{\beta}_j|)$. And we find a $\hat{\theta}$ that minimizes $GCV(\theta)$.

The third method based on BIC is easily constructed by functions defined above. The criterion is to select the optimal θ by minimizing

$$BIC(\theta) = \|\mathbf{Y} - \mathbf{X}\beta(\theta)\|^2 + p(\theta)\log(n)/n.$$

Wang et al. (2007), showed that the commonly used generalized cross-validation has an over-fitting effect in the resulting model for the SCAD variable selection procedure. Instead, they proposed that BIC is able to identify the true model consistently.

2.3.5 Asymptotic Properties

For the widespread use of LASSO in practice, it is important to know whether the LASSO selection is consistent. Here, consistent means that it correctly identify the variables for inclusion and exclusion from the model with probability increase to 1 as the sample size increases to infinite.

Meinshausen and Buhlmann (2004) showed that LASSO gives inconsistent variable selection results even with optimal tuning parameter from prediction criterion. Zou (2006) gave a necessary condition the underlying model must satisfy for the LASSO variable selection to be consistent. It can be concluded from those results that there are scenarios in which the LASSO selection cannot be consistent. Fan and Li (2000) studied a class of penalty functions including LASSO from a different perspective. They concluded that a penalty function satisfying both the conditions for sparsity and continuity must be singular at origin. We know that LASSO has these properties because L_1 penalty function is singular at the origin. But they suspect LASSO produces biased estimates because it shifts the resulting estimator by a constant t.

Zou(2006) proposed an improved version of LASSO: adaptive LASSO. Its estimates enjoy consistency, continuity and sparsity providing a proper choice of selection tuning parameter. The adaptive LASSO is defined as follows: Suppose that the ordinary least square estimate $\hat{\beta}$ exists. Define the weight vector $\hat{\mathbf{w}} = \frac{1}{|\hat{\beta}|^{\gamma}}$, where tuning parameter $\gamma > 0$. The adaptive LASSO estimates $\hat{\beta}^{*(n)}$ are given by

$$\hat{\beta}^{*(n)} = \arg\min_{\beta}\{\|\mathbf{Y} - \sum_{j=1}^{p} \mathbf{X}_j\beta_j\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j|\beta_j|\}.$$

With the efficient path algorithm for LASSO, LARS (Efron et al., 2004) or the shooting algorithm from Fu (1998), we can solve adaptive LASSO without extra loads in computations. Following Zou (2006), an adaptive LASSO solution can be found by the following algorithm.

Algorithm of Adaptive LASSO

1. Define $\mathbf{x}_j^{**} = \mathbf{x}_j / \hat{w}_j$, $j = 1, 2, \ldots, p$.

2. Solve the LASSO solutions for all λ_n,

$$\hat{\beta}^{**} = \arg\min_{\beta}\{\|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j^{**}\beta_j\|^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|\}.$$

3. Output $\hat{\beta}^{*(n)} = \hat{\beta}^{**} / \hat{w}_j$, $j = 1, 2, \ldots, p$.

2.4 Variable Selection in Non-linear Regression Models

Least Angle Regression is an efficient algorithm to solve LASSO problems in linear regression models. To solve variable selection problems in generalized linear models, Wang and Leng (2007) proposed a unified LASSO estimation process using the Least Square Approximation (LSA) so that objective functions for generalized linear models with LASSO type penalty can be transformed into their asymptotically equivalent least squares problems. The idea can be simply described as follows. Suppose the observed data $(\mathbf{X_i}, Y_i), i = 1 \ldots n$ are independent and identically distributed. Assume β is a parameter of interest and $L_n(\beta)$ is observed data log-likelihood function. Let $\tilde{\beta}$ be the maximum likelihood estimator (MLE). Hence, our objective function with LASSO penalty is given by

$$\frac{1}{n} L_n(\beta) - \lambda_n \sum_{j=1}^{p} |\beta_j|. \tag{2.9}$$

Using standard Taylor expansion for $L_n(\beta)$ at $\tilde{\beta}$, we have

$$\begin{aligned}\frac{1}{n}L_n(\beta) &\approx \frac{1}{n}L_n(\tilde{\beta}) + \frac{1}{n}\dot{L}_n(\tilde{\beta})^T(\beta-\tilde{\beta}) \\ &\quad + \frac{1}{2}(\beta-\tilde{\beta})^T\frac{L_n^{''}(\tilde{\beta})}{n}(\beta-\tilde{\beta}) \\ &= \frac{1}{n}L_n(\tilde{\beta}) + \frac{1}{2}(\beta-\tilde{\beta})^T\frac{L_n^{''}(\tilde{\beta})}{n}(\beta-\tilde{\beta}).\end{aligned}$$

Since $L_n^{'}(\tilde{\beta}) = 0$, the variable selection approximately minimizes

$$\arg\min\left[\frac{1}{2}(\beta-\tilde{\beta})^T\left\{-\frac{L_n^{''}(\tilde{\beta})}{n}\right\}(\beta-\tilde{\beta}) + \lambda_n\sum_{j=1}^{p}|\beta_j|\right].$$

Hence, the LARS algorithm can be easily applied. The adaptive LASSO may also be applied to the nonlinear regression.

Besides the method of unifying variable selection in generalized linear models with LARS algorithm using least square approximations, Park and Hastie (2007) introduced a path following algorithm for L_1 regularized generalized linear models. To solve (Equation 2.1) above, they proposed to use the predictor-corrector method to give the entire path of the coefficient estimates as tuning parameter λ changes. It starts with a maximum threshold of λ beyond which the only non-zero coefficient will be the intercept. As λ decreases, other variables enter into the active set. In each iteration of a potential factor joins the active set, it consists three steps: first determine the decrement in λ; then linearly approximate the corresponding change in the coefficients (predict step); and generate a new solution based on estimates from the predictor step (corrector step). After these steps, a test is perform for each variables outside the active set to check if it should join in. Repeat the corrector step until no more covariates are qualified to get in the active set.

The predictor-corrector algorithm gives the entire solution path for the coefficients β's with a varying λ and ensures the solutions are exact at the locations of λ where active set changes. It provides an alternative to using LARS algorithm to solve variable selection problems in generalized linear models.

2.5 Variable Selection via Smoothly Clipped Absolute Deviation Penalty

Fan and Li (2001) generalized the L_1 penalty to arbitrary function of L penalty functions and summarized three properties that a good penalty function should have in order for an estimator to be unbiasedness, sparsity and continuity. A new penalty function, SCAD, was proposed. The continuous differentiable penalty function is defined as follows.

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\}$$

for some $a > 2$ and $\theta > 0$. For the simple penalized least square problem Equation 2.6 with SCAD penalty, the solution is given by Fan (1997) as follows.

$$\hat{\theta} = \begin{cases} sign(z)(|z| - \lambda)_+, & \text{when } |z| \leq 2\lambda, \\ \{(a-1)z - sign(z)a\lambda\}/(a-2), & \text{when } 2\lambda < |z| \leq a\lambda, \\ z & \text{when } |z| > a\lambda. \end{cases}$$

We plot the thresholding rule for SCAD in Figure 3, from which we can see the solution for least square problem with SCAD penalty is unbiased for large estimated coefficients. With this penalty function, a form of penalized least squares for classical linear regression model is

$$\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + n\sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{2.10}$$

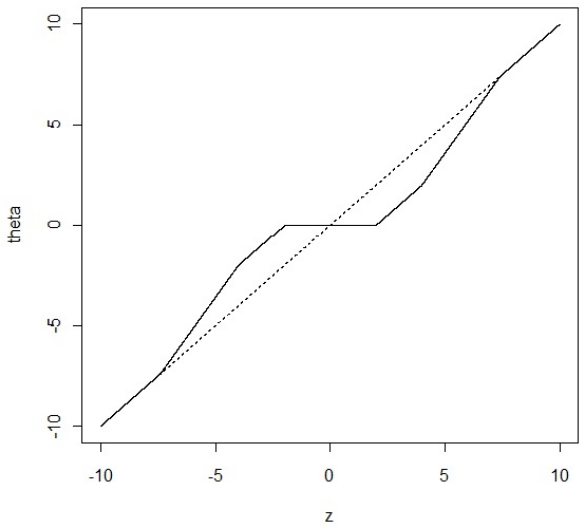

Figure 3. Plot of SCAD thresholding function with $\lambda = 2$ and $a = 3.7$

Since SCAD penalty functions are singular at the origin, they do not have continuous second order derivatives. A new local quadratic approximation algorithm was proposed in Fan and Li (2001). Suppose that we are given an initial value $\beta_{\mathbf{0}}$ that is close to the minimizer of (Equation 2.10) and β_{j0} is not very close to 0, we can use quadratic function to locally approximate SCAD function. That is

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\left\{\frac{p'_\lambda(|\beta_{j0}|)}{|\beta_{j0}|}\right\}(\beta_j^2 - \beta_{j0}^2)$$

for $\beta_j \approx \beta_{j0}$. Now, (Equation 2.10) is reduced to a quadratic minimization problem and the Newton-Raphson algorithm can be used. Specifically, the solution can be found by iteratively computing by

$$\beta_1 = \left\{\mathbf{X^T X} + n\Sigma_\lambda(\beta_\mathbf{0})\right\}^{-1} \mathbf{X}^T\mathbf{Y}$$

where $\Sigma_\lambda(\beta_\mathbf{0}) = diag\left\{\frac{p'_\lambda(|\beta_{10}|)}{|\beta_{10}|}, \ldots, \frac{p'_\lambda(|\beta_{p0}|)}{|\beta_{p0}|}\right\}$.

The weakness of this algorithm is its numerical instability. Fan and Li (2001) suggested that if any coefficient in a step of iteration is less than a prespecified value, i.e., very close to 0, then set it to zero and delete from the iteration. This extra process is adding another tuning parameter to estimate so it increases the computation load. To eliminate this weakness, Zou and Li (2008) proposed a new unified algorithm based on local linear approximation(LLA) to the penalty function:

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + p'_\lambda(|\beta_{j0}|)(|\beta_j| - |\beta_{j0}|)$$

for $\beta_j \approx \beta_{j0}$. From Fan and Chen (1999) and Cai, Fan and Li (2001), the one-step method is as efficient as the fully iterative method, provided that the initial estimators are reasonably good. Thus, one-step LLA estimator was proposed. The one-step LLA estimator possesses oracle properties. Furthermore, the LLA algorithm inherits the good features of LASSO in terms of computational efficiency. Therefore the one-step estimator can be solved by efficient algorithm for LASSO, such as the least angle regression(LARS) algorithm, described in Efron et al. (2004). For generalized linear models, denote $l(\beta) = \sum_{i=1}^n l_i(\beta)$ the model log-likelihood function. A SCAD penalized estimate can be obtained via solving

$$\hat{\beta} \quad = \quad \arg\max_\beta \left\{\sum_{i=1}^n l_i(\beta) - n\sum_{j=1}^p p_\lambda(|\beta_j|)\right\}$$

$$= \quad \arg\min_\beta \left\{ -\sum_{i=1}^{n} l_i(\beta) + n\sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}.$$

Suppose that the log-likelihood function is smooth and has the first two derivative with respect to β. For a given initial value $\beta^{(0)}$, the log-likelihood function can be locally approximated by

$$l(\beta) \approx l(\beta^{(0)}) + \nabla l(\beta^{(0)})^T(\beta - \beta^{(0)}) + \frac{1}{2}(\beta - \beta^{(0)})^T \nabla^2 l(\beta^{(0)})(\beta - \beta^{(0)}).$$

If we take $\beta^{(0)} = \hat{\beta}(mle)$, then $\nabla l(\beta^{(0)}) = 0$ by the definition of MLE. Thus, one-step estimate $\beta^{(1)}$ is given by

$$\beta^{(1)} = \arg\min_\beta \left\{ \frac{1}{2}(\beta - \beta^{(0)})^T[-\nabla^2 l(\beta^{(0)})](\beta - \beta^{(0)}) + n\sum_{j=1}^{p} p'_\lambda(|\beta_j^{(0)}|)|\beta_j| \right\}. \tag{2.11}$$

CHAPTER 3

VARIABLE SELECTION IN MISSING DATA PROBLEMS

3.1 Missing Data Problems

The problem of missing data is common in practice. Rubin (1976) and Little and Rubin (1987) classified missing data into three categories based on the missing data mechanisms. They are: missing completely at random (MCAR), where missingness does not depend on either observed or missing data; missing at random (MAR), where missingness only depends on observed data but not on missing data; and non-ignorable missing (NI), where missingness depends on unobserved data. There are three major categories of methods for missing data problems. The first one is the inverse missing data probability weighted approach. Since complete cases can not be regarded as a random sample from the population when the missing data mechanism is MAR, it is natural to use inverses of the missing data probability to weight the sample to correct bias when missing mechanism is MAR. This method is simple to apply but requires us to know the missing data probability.

The second method is the maximum likelihood (ML) approach. From Rubin (1976), when missing data is MAR and the conditional distribution of the missing data indicator given full data does not depend on parameters of interest, estimation and inference based on the observed data likelihood can be carried out without explicitly modeling the missing data mechanism when a parametric model is assumed for the complete data. But often maximizing the incompletely observed data likelihood is still a challenging task. Because maximizing the complete data likelihood is usually a much easier task than maximizing the incomplete data likelihood, Dempster, Laird, and Rubin (1977) formally introduced the EM algorithm

for maximizing the incomplete data likelihood through iteratively maximizing the expected complete data likelihood under the current estimated model. The general framework is as follows. Suppose $(Y_1, \ldots, Y_p) \sim f(Y_1, \ldots, Y_p, \theta)$ and the observed data are $\{R_i, R_i(Y_i)\}$. The complete data likelihood is

$$\prod_{i=1}^{n} f(Y_{i1}, \ldots, Y_{ip}, \theta).$$

Details of the EM algorithm are as follow.

EM Algorithm

1. Denote the initial value of parameter $\theta^{(0)}$.

2. E-step: Compute

$$Q(\theta|\theta^{(0)}) = \sum_{i=1}^{n} \sum_{r} E\left\{\log f(Y_{i1}, \ldots, Y_{ip}, \theta)|r_1(Y_{i1}), \ldots, r_p(Y_{ip}), \theta^{(0)}\right\}$$

3. M-step: Maximize $Q(\theta|\theta^{(0)})$ with respect to θ. Denote the maximizer by $\theta^{(1)}$.

4. Update $\theta^{(0)}$ by $\theta^{(1)}$ then repeat the process from step 2 until the current estimate is nearly unchanged from the previous estimate.

The advantage of EM algorithm is its computational convenience and easiness to implement for various missing data problems. The drawback of EM algorithm is that it in general converges slower than Newton-Raphson method. In addition, the E-step in EM algorithm may be hard to compute in nonlinear models. Thus, Monte Carlo and other numeric methods were proposed to solve these problems in the E-step.

The third method of handling missing data is imputation. Loosely speaking, the appeal of proper imputation is that we can somewhat treat imputed dataset as if they were fully observed when we perform data analysis (Schafer, 1999). In practice, most of the difficulties

lie in the creation of proper imputed values. Once imputed data set is obtained, inference may not be that difficult (Rubin, 1987; Rubin, 1996). Proper imputation often involves generating samples from a complex distribution for missing covariates conditional on observed covariates. Markov Chain Monte Carlo (MCMC) methods (Gelfand and Smith 1990) can be used to generate a draw from a distribution that approximates the complex distribution. One advantage of MCMC in computation is that increasing dimensionality usually does not slow convergence, which is an attractive feature in dealing with high dimension missing data problems. One of the common methods in MCMC is Gibbs sampler (Casella and George, 1992). A general illustration of Gibbs sampler in Bayesian imputation is as follows. Suppose that we want to generate missing values from the predictive distribution as

$$Y_1^{mis}, \ldots, Y_n^{mis} \sim p(y_1^{mis}, \ldots, y_n^{mis} | Y_1^{obs}, \ldots, Y_n^{obs}).$$

Given initial value $\theta^{(0)}$, the first step is to impute the missing values based on

$$Y_i^{mis} \sim f(y_i^{mis} | Y_i^{obs}, \theta^{(0)}).$$

Once Y_i^{mis}, $i = 1, \ldots, n$ are imputed, θ is updated by

$$\theta^{(1)} \propto \prod_{i=1}^{n} f(y_i^{mis} | Y_i^{obs}, \theta) p(\theta),$$

where $p(\theta)$ is the prior distribution. The above process is continued until convergence. In practice, Gibbs sampler may be paired with other sampling methods to generate a random draw from a targeted distribution.

3.2 Numerical Integration

In missing-data problems, we often need to compute complex integrations that usually do not have closed forms. Gaussian Quadrature (GQ) approximation is a powerful tool for evaluating these integrations. It improve the basic idea of numerical integrations that use equally spaced points as abscissas by choosing locations of these abscissas where functions can be exactly evaluated. With proper choices of abscissas and weights, GQ approximation is exact for a class of integrands which can be expressed as polynomials times some known function (Press et al., 2007). For an example, given a function $W(x)$ and an integer N, choosing a set of weights w_j and abscissas $x_j, j = 0, 1, \ldots, N-1$, we have following approximation

$$\int_a^b W(x) \times f(x)dx \approx \sum_{j=0}^{N-1} w_j \times f(x_j). \tag{3.1}$$

Furthermore, if $f(x)$ is a polynomial, the above approximation is exact. To find appropriate abscissas, we first need to find a set of normalized and mutually orthogonal polynomials (also called *orthonormal*). The abscissas we need in formula (Equation 3.1) with weighted function $W(x)$ are exactly the roots of these found orthonormal polynomials with respect for the same interval and weighting function. Hence, GQ procedure consists of three steps: 1) find a set of orthonormal polynomials with respect to weighting function; 2) solve these polynomials for their roots as abscissas, and 3) find the weights w_j. Though a GQ procedure sounds clear and easy to implement, computations of it can be quite difficult, depending on the weighting function. However, Press et al. (2007), provides individual subroutine programs that calculate the abscissas and weight for the most commonly used weight functions for their corresponding GQ formulas. Many GQ abscissas and weights involving classic weighting functions are also tabulated in books, such as Abramowitz and Stegun (1964) or Stroud and Secrest (1966). In

missing-data problems, the usual kind of integration we need is called Gauss-Hermite, in which weighting function is standard normal $W(x) = e^{-x^2}, -\infty < x < \infty$. It can be used to evaluate any integrands involving a function times a normal density. For an example, if $x \sim N(\mu, \sigma)$, the expectation of any known function $f(x)$ can be evaluated as follows.

$$\begin{aligned} E[f(x)] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\mu + \sigma y) e^{-\frac{y^2}{2}} dx \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} f(\mu + \sqrt{2}\sigma z) e^{-z^2} dx \\ &\approx \frac{1}{\sqrt{\pi}} \sum_{i=0}^{N-1} f(\mu + \sqrt{2}\sigma x_i) w_i, \end{aligned}$$

where $x_i, w_i, i = 1, \ldots, N-1$ are the abscissas and weights of the N-point Gauss-Hermite quadrature, respectively.

One disadvantage of GQ is that its complexity increases exponentially for multiple integrations. An alternative to GQ is Monte Carlo simulations, which in general are not as accurate as GQ but can be more efficient for high dimensional integrals. The basic idea of Monte Carlo is easy. We can rewrite the desired integrations as follows.

$$I \equiv \int_a^b g(x) f(x) dx.$$

Let $f(x)$ be the density function of X, then the right side of above equation is $E[g(x)]$. If we draw an *i.i.d* random sample $X_1, \ldots, X_n$ from $f(x)$, we can approximate I by the sample average and by law of large numbers:

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) \rightarrow_p I$$

as $n \to \infty$.

Compared with GQ, Monte Carlo simulation is easier to implement, but it converges slower at a speed of $O_p(n^{-1/2})$ (Givens and Hoeting, 2005). As dimension increases, computation remains almost the same in using the Monte Carlo approximation. On the other hand, GQ performs the best when dimension is small (Evans and Swartz, 2000). Hence, one needs to make a wise choice between these two numerical integration methods in practice to balance efficiency and accuracy, when computation load is acceptable.

3.3 Variable Selection with Missing Covariates

Variable selection is very challenging in missing data problems because the observed data likelihood often involves evaluation of multiple integrations that are not available in closed forms. Thus, Equation 2.11 is not directly available to use in implementing variable selection with SCAD penalty. Besides computational difficulties for the observed data log-likelihood, selecting appropriate penalty parameters to produce efficient estimates with suitable asymptotic properties such as sparsity and asymptotic normality is also challenging. The primary method of selecting penalty parameters use the five-fold cross validation method and generalized cross-validation (GCV) method. Wang and Leng (2007) and Wang, Li, and Tsai (2007) showed that in linear models, the GCV cannot identify the true model consistently but the BIC can.

3.3.1 Variable Selection Via Expectation-Maximization Algorithm

The EM algorithm (Dempster et al., 1977) maximizes the expected full data log-likelihood conditional on observed data (Q-function), reducing complexity of directly calculating the observed data likelihood. Ibrahim et al. (2008), proposed to maximize the penalized likelihood function, given by

$$l(\beta) - n\sum_{j=1}^{p} \phi_{\lambda_j}(|\beta_j|) = \sum_{i=1}^{n} l_i(\beta) - n\sum_{j=1}^{p} \phi_{\lambda_j}(|\beta_j|) \tag{3.2}$$

where λ_j is the penalty parameter corresponding to the jth regression coefficients. The penalty function $\phi_{\lambda_j}(\cdot)$ either takes the form of adaptive LASSO (Zou, 2006) or SCAD (Fan and Li, 2001). But because the observed-data log-likelihood function usually involves intractable integrations, they developed a Monte Carlo EM algorithm to compute the MPLE of β. Its E step is to evaluate the penalized Q-function

$$Q_\tau(\beta|\beta^{(s)}) = Q(\beta|\beta^{(s)}) - n\sum_{j=1}^{p}\phi_{\lambda_j}(|\beta_j|). \tag{3.3}$$

Since Q-function involves intractable integration, we can approximates it by taking a sequence of samples from Gibbs Sampler (Geman and Geman 1984) along with the adaptive rejection algorithm of Gilks and Wild (1992); then uses the Monte Carlo version of the EM algorithm given by Wei and Tanner (1990). Implement details are described in Ibrahim et al. (2008).

In general, usual criteria for selection of penalty parameters for missing-data problems including the five-fold cross validation, GCV and BIC cannot be easily computed. Ibrahim et al. (2008), proposed two methods to select penalty parameter: an IC_Q criterion that selects optimal λ by minimizing

$$IC_Q(\lambda) = -2Q(\hat{\beta}_\lambda|\hat{\beta}_0) + \dim(\beta) \times \log(n);$$

and an $IC_{H,Q}$ that only uses observed data likelihood

$$IC_{H,Q}(\lambda) = -2\log f(x_{obs}|\hat{\beta}_0) + \dim(\beta) \times \log(n).$$

where $\hat{\beta}_0$ is the unpenalized estimator from EM algorithm. They showed in their paper under certain regularity conditions, the maximized penalized likelihood estimator has the Oracle properties based on the $IC_{H,Q}$ selection criterion.

3.3.2 Variable Selection via Imputation

Since computation in likelihood approach can be difficult in variable selection with missing covariates, alternative approaches are explored. Yang et al. (2005), proposed to use a Bayesian stochastic search approach combined with multiple imputation. It considers a multivariate normal case with p independent variables, $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_p}$ and a dependent variable $\mathbf{Y}$. The model they used for variable selection is

$$\mathbf{Y} = \alpha\mathbf{1} + \sum_{j=1}^{p} \gamma_i \mathbf{X_j} \beta_j + \epsilon, \epsilon \sim N_n(0, \sigma^2 \mathbf{I})$$

where the indicator $\gamma_j = 1$ or $\gamma_j = 0$ corresponds to the inclusion or exclusion of $\mathbf{X_j}$, respectively. Two approaches are proposed. One is called "impute, then select" (ITS). It adopts the following hierarchical prior distribution from George and McCulloch (1993):

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2)$$

$$\sigma^2|\gamma \sim IG(\frac{v}{2}, \frac{v\lambda_v}{2})$$

and

$$p(\gamma) \sim \prod_{j=1}^{p} w_j^{\gamma_j}(1-w_j)^{1-\gamma_j}$$

where $w_j = p(\gamma_j = 1)$ and c_j, τ_j, v and λ_γ are constants. George and McCulloch (1993) gives details of specific choices of them. To implement ITS, we first impute m data sets, $\mathbf{D^{(1)}}, \mathbf{D^{(2)}}, \ldots, \mathbf{D^{(m)}}$. Then perform data analysis for each imputed data set to obtain parameter

estimates $(\beta_\gamma^{(l)}, \sigma^{(l)})$ and their variance estimates. Multiple imputation combining rules (Rubin 1987) can be applied to synthesize these m sets of results into a final summary.

The second method is called "Simultaneously Impute and Select". It uses the linkage between Schafer's imputation algorithm (Schafer, 1997) and George and McCulloch's variable selection algorithm and a one-to-one relationship between parameters of joint model and parameters in the partitioned representation: i.e., suppose we have $\mathbf{X} \sim N(\mu_x, \Sigma_x)$, $\mathbf{Y}|\mathbf{X} \sim N(\alpha\mathbf{1} + \mathbf{X}\beta, \sigma^2\mathbf{I})$ and partition$(\mu, \mathbf{\Sigma})$ into

$$\left(\begin{bmatrix} \mu_y \\ \mu_\mathbf{x} \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_x \end{bmatrix} \right)$$

then we have

$$\mu_y = \alpha + \mu_x\beta$$

$$\Sigma_{xy} = \Sigma_x\beta$$

$$\sigma_y^2 = \sigma^2 + \Sigma_{xy}\Sigma_x^{-1}\Sigma_{yx}.$$

It implements in the following steps: first impute the missing data; then generate parameters $\mu_\mathbf{x}, \mathbf{\Sigma_x}$ from the complete covariate matrix; using Bayesian variable selection to draw model parameters $\theta^{(\mathbf{t+1})} = (\gamma^{(t+1)}, \alpha^{(t+1)}, \sigma^{2(t+1)}, \beta^{(t+1)})$, see George and McCulloch (1993); lastly derive $\varphi^{(\mathbf{t+1})} = (\gamma, \mu_y, \sigma_y^2, \Sigma_{xy}, \mu_\mathbf{x}, \mathbf{\Sigma_x})$ from $\theta^{(\mathbf{t+1})}$ above. The disadvantage of imputation method in variable selection is though it is easy to implement, its asymptotic properties are unclear.

3.4 Problems with Existing Variable Selection Approaches

Fan and Li (2001) established the asymptotic theory for the non-concave penalized likelihood estimator for linear regression models and generalized linear models, but it is hard

to apply it to the missing data problems due to complicated second derivatives in missing data likelihood. Ibrahim et al. (2008), proposed a unified model selection and estimation procedure based on iteratively maximizing the penalized Q-function of an EM algorithm. The Q-function is not a good approximation to the observed data log-likelihood; in addition, they only consider model selection for the regression model for dependent variable but not for any covariates.

Yang et al. (2005), used Bayesian framework with MCMC sampling and multiple imputation into linear regression models but its asymptotic properties are unknown. For tuning parameter selection, Ibrahim et al. (2008), proposed two model selection criteria for general missing data problems. One is the IC_Q based only on conditional expectation with respect to observed data, and the other is based on the observed data log-likelihood, more like BIC for model selection for complete data problems but it is computationally expensive to implement.

Furthermore, current models for variable selection for data with missing data all focus on selecting covariates for the dependent variables. When we have many covariates subject to missing value, we need to fit a more comprehensive model to the data so that selection of variables for the covariate models becomes necessary.

CHAPTER 4

VARIABLE SELECTION USING EXPECTATION-MAXIMIZATION ALGORITHM

4.1 Expectation-Maximization Algorithm for Unpenalized Log-Likelihood

Let $Y = (y_1, \cdots, y_p)$ denote the full data vector from a subject. Suppose that the fully observed data are $Y_1, \cdots, Y_n$ which are independent and identically distributed. Let $f(y \mid \theta)$ denote the joint density function for Y. The likelihood for the fully observed data is

$$L^F(\theta \mid Y_1, \cdots, Y_n) = \prod_i^n f(Y_i \mid \theta). \tag{4.1}$$

Suppose further that Y is subject to missing value and the missing values are missing at random in Rubin's (1976) sense. Let Y^{mis} and Y^{obs} respectively denote the missing part and the observed part of Y. The likelihood for the observed data is

$$L(\theta \mid Y_1^{obs}, \cdots, Y_n^{obs}) = \prod_i^n g(Y_i^{obs} \mid \theta), \tag{4.2}$$

where

$$g(Y^{obs} \mid \theta) = \int f(Y^{mis}, Y^{obs} \mid \theta) dY^{mis}.$$

The observed data likelihood is usually intractable to work with and the full data likelihood is often much easier to handle. The EM algorithm (Dempster et al., 1977; Little and Rubin, 2002) was introduced to maximize the observed data log-likelihood through iteratively maximizing the expected full data log-likelihood given the observed data.

Define the expected full data log-likelihood given the observed data under the current estimated model by

$$Q(\theta \mid \theta^*) = E\left\{\log f(Y^{mis}, Y^{obs} \mid \theta) \mid Y^{obs}, \theta^*\right\}, \tag{4.3}$$

where the expectation E is taken under the current model

$$h(Y^{mis} \mid Y^{obs}, \theta^*) = \frac{f(Y \mid \theta^*)}{g(Y^{obs} \mid \theta^*)}.$$

Following Ibrahim et al. (2008), define

$$H(\theta \mid \theta^*) = E\left\{\log h(Y^{mis} \mid Y^{obs}, \theta) \mid Y^{obs}, \theta^*\right\}.$$

It is now well-known that

$$l(\theta \mid Y_1^{obs}, \cdots, Y_n^{obs}) = Q(\theta \mid \theta^*) - H(\theta \mid \theta^*), \tag{4.4}$$

where $l(\theta \mid Y_1^{obs}, \cdots, Y_n^{obs}) = \log L(\theta \mid Y_1^{obs}, \cdots, Y_n^{obs})$. By Jensen's inequality for convex function,

$$H(\theta \mid \theta^*) \leq H(\theta^* \mid \theta^*). \tag{4.5}$$

When there exists an θ^{**} such that

$$Q(\theta^{**} \mid \theta^*) \geq Q(\theta^* \mid \theta^*), \tag{4.6}$$

it follows that

$$l(\theta^{**} \mid Y_1^{obs}, \cdots, Y_n^{obs}) \geq l(\theta^{*} \mid Y_1^{obs}, \cdots, Y_n^{obs}).$$

The EM algorithm uses this idea to maximize the observed data log-likelihood $l(\theta \mid Y_1^{obs}, \cdots, Y_n^{obs})$ by iteratively maximizing $Q(\theta \mid \theta^*)$ with θ^* being updated to the current estimate of θ in each iteration.

4.2 Expectation-Maximization Algorithm for Penalized Likelihood

The model selection with incompletely observed data usually maximizes the penalized observed data likelihood as

$$l_\lambda(\theta \mid Y_1^{obs}, \cdots, Y_n^{obs}) = l(\theta \mid Y_1^{obs}, \cdots, Y_n^{obs}) - n\sum_{j=1}^{K} p_{j\lambda}(|\theta_j|), \tag{4.7}$$

where θ is a vector with individual elements $\theta_1, \cdots, \theta_K$. As in the case of maximizing the unpenalized observed data likelihood, we still face the problem that the observed data log-likelihood can be intractable and difficult to work with in the maximization. Analogue to the EM algorithm for the unpenalized log-likelihood, an EM algorithm for the penalized likelihood can be carried out as follows. Define a penalized analogue to the Q-function in the EM algorithm as

$$Q_\lambda(\theta \mid \theta^*) = Q(\theta \mid \theta^*) - n\sum_{j=1}^{K} p_{j\lambda}(|\theta_j|). \tag{4.8}$$

The penalized log-likelihood for the incompletely observed data can be expressed as

$$l_\lambda(\theta \mid Y_1^{obs}, \cdots, Y_n^{obs}) = Q_\lambda(\theta \mid \theta^*) - H(\theta \mid \theta^*). \tag{4.9}$$

By following the same arguments in designing the EM algorithm for the unpenalized likelihood, an EM algorithm can be carried out by maximizing the penalized $Q_\lambda(\theta \mid \theta^*)$ repeatedly. The

E-step for the EM algorithm for the penalized likelihood is the same as that for the unpenalized likelihood. The M-step of the EM algorithm for the penalized likelihood maximizes $Q_\lambda(\theta \mid \theta^*)$. If there exists a θ^{**} such that

$$Q_\lambda(\theta^{**} \mid \theta^*) \geq Q_\lambda(\theta^* \mid \theta^*), \tag{4.10}$$

then it follows that

$$l_\lambda(\theta^{**} \mid Y_1^{obs}, \cdots, Y_n^{obs}) \geq l_\lambda(\theta^* \mid Y_1^{obs}, \cdots, Y_n^{obs}). \tag{4.11}$$

This means that the EM algorithm for the penalized log-likelihood also increases the log-likelihood in each iteration.

4.3 Algorithm for Maximizing $Q_\lambda(\theta \mid \theta^*)$

The Q-function in the EM algorithm can be approximated by the following quadratic form

$$\begin{aligned} Q(\theta \mid \theta^*) &\approx Q(\theta^* \mid \theta^*) + (\theta - \theta^*)^T \dot{Q}(\theta^* \mid \theta^*) + \frac{1}{2}(\theta - \theta^*)^T \ddot{Q}(\theta^* \mid \theta^*)(\theta - \theta^*) \\ &= Q(\theta^* \mid \theta^*) + \frac{1}{2}\dot{Q}^T(\theta^* \mid \theta^*)\{\ddot{Q}(\theta^* \mid \theta^*)\}^{-1}\dot{Q}(\theta^* \mid \theta^*) \\ &\quad + \frac{1}{2}(\theta - \theta^* - d^*)^T \ddot{Q}(\theta^* \mid \theta^*)(\theta - \theta^* - d^*), \end{aligned}$$

where $d^* = -\{\ddot{Q}(\theta^* \mid \theta^*)\}^{-1}\dot{Q}(\theta^* \mid \theta^*)$. Maximizing the penalized Q-function can be approximately performed by maximizing

$$-\frac{1}{2}(\theta - \theta^* - d^*)^T\{-\ddot{Q}(\theta^* \mid \theta^*)\}(\theta - \theta^* - d^*) - n\sum_{j=1}^{K} p_{j\lambda}(|\theta_j|). \tag{4.12}$$

When $p_{j\lambda}(|\theta_j|) = \lambda|\theta_j|$, the foregoing maximization problem becomes the penalized least square with L^1-penality when $\{-\ddot{Q}(\theta^* \mid \theta^*)\}$ is non-negative definite. The LARS algorithm or the coordinate descent algorithm can be applied. For a general penalty function such as SCAD, linear approximation may be applied to reduce the problem into a problem with an L^1 penalty.

In many applications, θ can be split into two parts. One part denoted by η is subject to penalty and the other part denoted by γ is not subject to penalty. The penalized Q-function becomes

$$Q_\lambda(\eta, \gamma \mid \eta^*, \gamma^*) = Q(\eta, \gamma \mid \eta^*, \gamma^*) - n\sum_{j=1}^{k} p_{j\lambda}(|\eta_j|), \tag{4.13}$$

where k is the number of element in η. By a similar quadratic expansion, the maximization problem is equivalent to the least square minimization problem as

$$\frac{1}{2}(\eta - \eta^* - d_\eta^*, \gamma - \gamma^* - d_\gamma^*)^T\{-\ddot{Q}(\theta^* \mid \theta^*)\}(\eta - \eta^* - d_\eta^*, \gamma - \gamma^* - d_\gamma^*) + n\sum_{j=1}^{k} p_{j\lambda}(|\eta_j|), \tag{4.14}$$

where $d^* = (d_\eta^*, d_\gamma^*)$. The coordinate descent algorithm can be applied directly. For the application of the LARS algorithm, the M-step of the EM algorithm can be further divided into two steps, one maximizes the penalized Q-function with respect to η with γ fixed and the other maximizes Q with respect to γ with η fixed. The former can be done using LARS algorithm and the latter can be done simply using the Newton-Raphson algorithm or in some cases, a closed-form solution.

4.4 One Step Algorithm for Maximizing the Penalized Likelihood

When the number of the variables is much smaller than the sample size, the unpenalized maximum likelihood estimator $\hat{\theta}$ is consistent and symptotically normally distributed. The penalized log-likelihood can be carried out starting from the maximum unpenalized likelihood

estimator. In this case, the maximum penalized likelihood approach is asymptotically equivalent to minimizing

$$\frac{1}{2}(\theta-\hat{\theta})^T\{-\ddot{l}(\hat{\theta})\}(\theta-\hat{\theta})+n\sum_{j=1}^{K}p_{j\lambda}(|\theta_j|). \tag{4.15}$$

Note from the Fisher's information identity that

$$-\ddot{l}(\hat{\theta})\approx\sum_{i=1}^{n}\dot{l}_i^T(\hat{\theta})\dot{l}_i(\hat{\theta}), \tag{4.16}$$

where $l_i=\log g(Y_i^{obs}\mid\hat{\theta})$. As a by-product of the EM algorithm, it can be seen that

$$\dot{l}_i(\hat{\theta})=\dot{Q}_i(\hat{\theta}\mid\hat{\theta}), \tag{4.17}$$

where $Q_i(\theta\mid\theta^*)=E\{\log f(Y_i\mid\theta)\mid Y_i^{obs},\theta^*\}$. The penalized maximum likelihood estimator can be obtained by minimizing

$$\frac{1}{2}(\theta-\hat{\theta})^T\left\{\sum_{i=1}^{n}\dot{Q}_i^T(\hat{\theta}\mid\hat{\theta})\dot{Q}_i(\hat{\theta}\mid\hat{\theta})\right\}(\theta-\hat{\theta})+n\sum_{j=1}^{K}p_{j\lambda}(|\theta_j|). \tag{4.18}$$

4.5 Selection of the Tuning Parameter

Suppose that, for each fixed λ, we find a maximum penalized likelihood estimator, denoted by θ_λ. Many approaches for selecting the tuning parameter λ are based on the observed data likelihood which may not be readily available. For example, in the BIC selection of tuning parameter, the following function

$$2l(\theta_\lambda)-K\log(n) \tag{4.19}$$

is minimized with respect to λ, where K is the number of parameters in the model and n is the sample size. Note that $l(\theta)=Q(\theta\mid\theta^*)-H(\theta\mid\theta^*)$. In the EM algorithm, Q-function is a

by product of the algorithm. However, H function is unavailable to use. Ibrahim et al. (2008), proposed a fairly complicated approximation based on mixture of normals to approximate the conditional distribution of the missing data given the observed data and relied on Monte Carlo sample to approximate the H function. To alleviate the computational problem, Ibrahim et al. (2008), also proposed a criterion called IC_Q that uses Q-function in place of the log-likelihood function in the selection of the tuning parameter. Although Garcia et al. (2010), and Ibrahim et al. (2011), showed that the their IC_Q criterion can consistently select the important covariates. Nevertheless, such a tuning parameter selection approach asymptotically selects more variables then necessary into the model. As a result, false positive or over-fit often occurs.

This problem can be resolved in the following way. Note that

$$l(\theta_\lambda) \approx l(\hat{\theta}) + (\theta_\lambda - \hat{\theta})^T \dot{l}(\hat{\theta}) + \frac{1}{2}(\theta_\lambda - \hat{\theta})^T \ddot{l}(\hat{\theta})(\theta_\lambda - \hat{\theta}), \tag{4.20}$$

where $\hat{\theta}$ is the maximum likelihood estimator. Since $\dot{l}(\hat{\theta}) = 0$, it follows that

$$l(\theta_\lambda) - l(\hat{\theta}) \approx \frac{1}{2}(\theta_\lambda - \hat{\theta})^T \ddot{l}(\hat{\theta})(\theta_\lambda - \hat{\theta}), \tag{4.21}$$

By the same arguments as in finding the penalized maximum likelihood estimator, the tuning parameter can be selected using the approximation to the observed data log-likelihood as

$$\frac{1}{2}(\theta_\lambda - \hat{\theta})^T \left\{ \sum_{i=1}^{n} \dot{Q}_i^T(\hat{\theta} \mid \hat{\theta}) \dot{Q}_i(\hat{\theta} \mid \hat{\theta}) \right\} (\theta_\lambda - \hat{\theta}) + c(n, K), \tag{4.22}$$

where $c(n, K)$ is a penalty term. In the BIC selection criterion, $c(n, K) = K \log n$. One advantage of the proposed approximation approach over the approximation by Q-function proposed in Ibrahim et al (2008) is that the proposed approach is asymptotically equivalent to

the use of the observed data log-likelihood. In contrast, the IC_Q approach proposed in Ibrahim et al. (2008), and used in Garcia et al. (2010), and Ibrahim et al. (2011), is not.

To carry out the proposed approximation, we need to run several EM algorithms. One is for the unpenalized log-likelihood and the others are for the penalized log-likelihood with different λ. The maximum likelihood estimators are respectively denoted by $\hat{\theta}$ and θ_λ for different λ.

4.6 Some Theoretical Results

Suppose we have data $\mathbf{Y_1}, \ldots, \mathbf{Y_n} \sim_{iid} N(\mu, \Sigma)$. We can write the model in a consecutive regression models format as follows.

$Y_1|Y_2, \ldots, Y_p \sim N(\eta_1, \sigma_1^2)$

$Y_2|Y_3, \ldots, Y_p \sim N(\eta_2, \sigma_2^2)$

...

$Y_{p-1}|Y_p \sim N(\eta_{p-1}, \sigma_{p-1}^2)$

$Y_p \sim N(\eta_p, \sigma_p^2)$,

where $\eta_j = \beta_{j0} + \beta_{j,(j+1)}Y_{j+1} + \ldots + \beta_{j,p}Y_p,\ j = 1, \ldots, (p-1)$. Let $\theta = (\gamma, \eta)$, where $\gamma = (\beta_{10}, \ldots, \beta_{(p-1)0}, \sigma_1^2, \ldots, \sigma_{p-1}^2)$ and $\eta = (\beta_{1,2}, \ldots, \beta_{1,p}, \beta_{2,3}, \ldots, \beta_{(p-1)p})$. There is then a one-to-one relationship between parameters in (μ, Σ) and parameters γ and η. The objective penalized function for the model selection problem is

$$\begin{aligned} l_\lambda(\eta) &= \sum_{i=1}^{n} \log g(\mathbf{Y}^{obs}, \gamma, \eta) - n\sum_{j=1}^{p'} p_{\lambda_n}(|\eta_j|) \\ &= l(\mathbf{Y}^{obs}, \eta, \gamma) - n\sum_{j=1}^{p'} p_{\lambda_n}(|\eta_j|) \end{aligned} \tag{4.23}$$

where $p' = \frac{p(p-1)}{2}$ is the number of consecutive regression coefficients and $g(\mathbf{Y}^{obs}, \theta)$ is the marginal likelihood for the observed data. Following Fan and Li (2001), we prove the existence

and consistency of maximum penalized likelihood estimator.

Theorem 1. Let $\mathbf{Y_1}, \ldots, \mathbf{Y_n}$ be independent and identically distributed with density $f(\mathbf{X}, \eta, \gamma)$ that is multivariate normal $N(\mu, \Sigma)$. If $\max\{|p''_{\lambda_n}(|\eta_{j0}|)| : \eta_{j0} \neq 0\} \to 0$, then there exists a local maximizer $\hat{\eta}$ of $l_\lambda(\eta)$ such that $\|\hat{\eta} - \eta_0\| = O_p(n^{-1/2} + a_n)$ where $a_n = \max\{p'_{\lambda_n}(|\eta_{j0}|) : \eta_{j0} \neq 0\}$.

Proof: Let $\alpha_n = n^{-1/2} + a_n$. We want to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\{\sup_{\|\mathbf{u}\|=C} l_{\lambda_n}(\eta_0 + \alpha_n \mathbf{u}) < l_{\lambda_n}(\eta)\} \geq 1 - \varepsilon. \tag{4.24}$$

It implies probability goes to 1 as $n \to \infty$ that there exists a local maximum in the ball $\{\eta_0 + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$ and this maximum has the property $\|\hat{\eta} - \eta_0\| = O_p(\alpha_n)$.

$$\begin{aligned}
D_n(\mathbf{u}) &\equiv Q(\eta_0 + \alpha_n \mathbf{u}) - Q(\eta_0) \\
&= L(\mathbf{Y}^{obs}, \eta_0 + \alpha_n \mathbf{u}) - L(\mathbf{Y}^{obs}, \eta_0) \\
&\quad - n \sum_{j=1}^{p'} \{p_{\lambda_n}(|\eta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\eta_{j0}|)\} \\
&\leq L(\mathbf{Y}^{obs}, \eta_0 + \alpha_n \mathbf{u}) - L(\mathbf{Y}^{obs}, \eta_0) \\
&\quad - n \sum_{j=1}^{s} \{p_{\lambda_n}(|\eta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\eta_{j0}|)\} \\
&= \alpha_n L'(\mathbf{Y}^{obs}, \eta_0)^T \mathbf{u} - \frac{1}{2} n \alpha_n^2 \mathbf{u}^T I(\mathbf{Y}^{obs}, \eta_0) \mathbf{u}\{1 + o_p(1)\} - \\
&\quad n \sum_{j=1}^{s} [\alpha_n p'_{\lambda_n}(|\eta_{j0}|) \mathrm{sgn}(\eta_{j0}) u_j + \alpha_n^2 p''_{\lambda_n}(|\eta_{j0}|) u_j^2 \{1 + o_p(1)\}]
\end{aligned} \tag{4.25}$$

where s is the number of non-zero components of η and the second inequality comes from that $p_{\lambda_n}(0) = 0$ for SCAD penalty function. Since $n^{-1/2} L'(\mathbf{Y}^{obs}, \eta_0) = O_p(1)$, the first term from

last equality is of the order $O_p(n^{1/2}\alpha_n) = O_p(n\alpha_n^2)$. Thus, the second term dominates the first term by picking a sufficient large C. And the third term is bounded by

$$sn\alpha_n a_n\|\mathbf{u}\| + n\alpha_n^2 \max\{|p''_{\lambda_n}(|\eta_{j0}|)| : \eta_{j0} \neq 0\}\|\mathbf{u}\|^2$$

that is also dominated by the second term. Thus, by choosing a large enough constant C, equation (Equation 4.24) holds.

To see $n^{-1/2}L'(\mathbf{Y}^{obs}, \eta_0) = O_p(1)$, we show that $n^{-1/2}L'(\mathbf{Y}^{obs}, \eta_0) \sim N(0, I^{-1}(\eta_0))$. First, we write

$$\frac{\partial}{\partial\eta}\log f(Y_i^{obs}, \eta_0) = \sum_k 1_{(R_i=r_k)}\frac{\partial}{\partial\eta}\log f(r_k(Y_i), \eta_0).$$

We have

$$\begin{aligned}
E_\eta\left[\frac{\partial}{\partial\eta}\log f(Y_i^{obs}, \eta_0)\right] &= E\left[\sum_k 1_{(R_i=r_k)}\frac{\partial}{\partial\eta}\log f(r_k(Y_i), \eta_0)\right] \\
&= \sum_k E\left[\frac{\partial}{\partial\eta}\log f(r_k(Y_i), \eta_0)P(R_i = r_k|r_k(Y_i))\right] \\
&= \sum_k \int_R \frac{\partial}{\partial\eta}\log f(r_k(Y_i), \eta_0)P(R_i = r_k|r_k(Y_i))f(r_k(Y_i), \eta_0)dr_k(Y_i) \\
&= \sum_k \int_R \frac{\partial}{\partial\eta} f(r_k(Y_i), \eta_0)P(R_i = r_k|r_k(Y_i))dr_k(Y_i) \\
&= \frac{\partial}{\partial\eta}\sum_k \int_R f(r_k(Y_i), \eta_0)P(R_i = r_k|r_k(Y_i))dr_k(Y_i) \\
&= \frac{\partial}{\partial\eta}\sum_k P(R_i = r_k|\eta_0) \\
&= 0. \qquad (4.26)
\end{aligned}$$

And by Louis' formula, the observed information matrix for incompletely observed data is given by

$$I(\mathbf{Y}^{obs}, \eta_0) = E_\eta\left[-\frac{\partial^2}{\partial\eta^2}\log f(Y|\eta)|r(\mathbf{Y}), \eta_0\right] + E_\eta\left[(\frac{\partial}{\partial\eta}\log f(Y|\eta))^2|r(\mathbf{Y}), \eta_0\right]$$

Hence from central limit theorem (CLT), we have

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\eta}\log f(Y_i^{obs},\eta_0) = n^{-1/2}L^{'}(\mathbf{Y}^{obs},\eta_0) \sim N(0, I^{-1}(\eta_0))$$

and $n^{-1/2}L^{'}(\mathbf{Y}^{obs},\eta_0) = O_p(1)$, where $I(\eta_0) = E\left[I(\mathbf{Y}^{obs},\eta_0)\right]$.

Theorem 2. Let $\mathbf{Y_1},\ldots,\mathbf{Y_n}$ be independent and identically distributed with density $f(\mathbf{Y},\eta)$ that is multivariate normal $N(\mu,\Sigma)$. Assume that

$$\lim_{n\to\infty}\inf\lim_{\theta\to 0_+}\inf\frac{p_{\lambda_n}^{'}(\theta)}{\lambda_n} > 0.$$

If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n\to\infty$, then with probability tending to 1, for any given $\eta_\mathbf{1}$ satisfying $\|\eta_1 - \eta_{10}\| = O_p(n^{-1/2})$ and any constant C,

$$Q\begin{pmatrix}\eta_1\\ \mathbf{0}\end{pmatrix} = \max_{\|\eta_2\|\le Cn^{-1/2}} Q\begin{pmatrix}\eta_1\\ \eta_\mathbf{2}\end{pmatrix} \tag{4.27}$$

Proof: To show (Equation 4.27), it is sufficient to show that with probability tending to 1 as $n\to\infty$, for any η_1 satisfying $\|\eta_1 - \eta_{10}\| = O_p(n^{-1/2})$ and for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1,\ldots,p^{'}$,

$$\frac{\partial Q(\eta)}{\partial\eta_j} = \begin{cases} <0 & \text{for } 0<\eta_j<\varepsilon_n,\\ >0 & \text{for } -\varepsilon<\eta_j<0.\end{cases} \tag{4.28}$$

Using Taylor's expansion, we have

$$\frac{\partial Q(\eta)}{\partial\eta_j} = \frac{\partial L(\mathbf{Y}^{obs},\eta)}{\partial\eta_j} - np_{\lambda_n}^{'}(|\eta_j|)\mathrm{sgn}(\eta_j)$$

$$
\begin{aligned}
= \quad & \frac{\partial L(\mathbf{Y}^{obs}, \eta_0)}{\partial \eta_j} + \sum_{l=1}^{p'} \frac{\partial^2 L(\mathbf{Y}^{obs}, \eta_0)}{\partial \eta_j \partial \eta_l}(\eta_l - \eta_{l0}) \\
& + \sum_{l=1}^{p'} \sum_{k=1}^{p'} \frac{\partial^3 L(\mathbf{Y}^{obs}, \eta^*)}{\partial \eta_j \partial \eta_l \partial \eta_k}(\eta_l - \eta_{l0})(\eta_k - \eta_{k0}) \\
& - n p_{\lambda_n}^{'}(|\eta_j|)\mathrm{sgn}(\eta_j)
\end{aligned}
$$

where η^* lies between η and η_0. From previous statements, we know $n^{-1/2} L^{'}(\mathbf{Y}^{obs}, \eta_0) = O_p(1)$. So $L^{'}(\mathbf{Y}^{obs}, \eta_0) = O_p(n^{1/2})$. Since the observe information matrix valued at η_0 is finite,

$$
\frac{1}{n} \frac{\partial^2 L(\mathbf{Y}^{obs}, \eta_0)}{\partial \eta_j \partial \eta_l} = E\left[\frac{\partial^2 L(\mathbf{Y}^{obs}, \eta_0)}{\partial \eta_j \partial \eta_l} \right] + o_p(1),
$$

and the third derivatives are bounded in normality assumption, the first three terms above are all $O_p(n^{1/2})$. Thus, we have

$$
\begin{aligned}
\frac{\partial Q(\eta)}{\partial \eta_j} &= O_p(n^{1/2}) - n p_{\lambda_n}^{'}(|\eta_j|)\mathrm{sgn}(\eta_j) \\
&= n\lambda_n \{ O_p(n^{-1/2}/\lambda_n) - \lambda_n^{-1} p_{\lambda_n}^{'}(|\eta_j|)\mathrm{sgn}(\eta_j) \}.
\end{aligned}
$$

When $n^{-1/2}\lambda_n \to 0$ and $\liminf_{n\to\infty} \liminf_{\theta \to 0+} \frac{p_{\lambda_n}^{'}(\theta)}{\lambda_n} > 0$, the sign of this derivative is completely determined by sign of η_j. Hence, Equation 4.28 follows.

CHAPTER 5

VARIABLE SELECTION IN LINEAR REGRESSION WITH MISSING COVARIATES

5.1 Problem Description: A Simple Case

In this section, we show that the complexity of deriving the maximum observed data log-likelihood, even for a simple case where there is only one covariate, is subject to missing values. Therefore, we will instead use EM algorithm in the following section to maximize the observed data log-likelihood. Suppose we have a linear regression model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon, \tag{5.1}$$

where $\varepsilon \sim N(0, \sigma^2)$ and $\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$ is an $n \times p$ matrix. For simplicity of description, we first consider the case where only X_1 is subject to missing and the missing values are MAR. We can write the likelihood as follows.

$$\begin{aligned} L &= \prod_{i=1}^{n} \{f(y_i, x_{i1}|x_{i2}, ..., x_{ip})\}^{R_i} \{g(y_i|x_{i2}, \ldots, x_{ip})\}^{1-R_i} \\ &= \prod_{i=1}^{n} \{f_1(y_i|x_{i2}, \ldots, x_{ip}) f_2(x_{i1}|x_{i2}, ..., x_{ip})\}^{R_i} \\ &\quad \left\{\int f_1(y_i|x_1, ..., x_{pi}) \cdot f_2(x_1|x_{2i}, ..., x_{pi}) dx_1\right\}^{1-R_i}. \end{aligned} \tag{5.2}$$

Assume that $x_1|x_2, \ldots, x_p \sim N(\alpha_0 + \sum_{j=2}^{p} \alpha_j x_j, \tau^2)$. The conditional densities are

$$f_1(y_i|x_{i1}, \cdots, x_{ip}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2\right\}, \tag{5.3}$$

and

$$f_2(x_{i1}|x_{i2},\cdots,x_{ip}) = \frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{1}{2\tau^2}(x_{i1}-\alpha_0-\alpha_1x_{2i}-\cdots-\alpha_{p-1}x_{ip})^2\right\}. \tag{5.4}$$

Because $(y, x_1|x_2, ..., x_p)$ is bivariate normal, $y|x_2, ..., x_p$ is normally distributed with mean and variance as follows.

$$\begin{aligned}
E(y|x_2,...,x_p) &= E\left[E(y|x_1,...,x_p)|x_2,...,x_p\right] \\
&= E\left(\beta_0+\beta_1x_1+...+\beta_px_p|x_2,..,x_p\right) \\
&= \beta_0+\beta_1E(x_1|x_2,...,x_p)+\beta_2x_2+...+\beta_px_p \\
&= \beta_0+\beta_1(\alpha_0+\alpha_1x_2+...+\alpha_px_p)+\beta_2x_2+...+\beta_px_p \\
&= (\beta_0+\beta_1\alpha_0)+(\beta_1\alpha_1+\beta_2)x_2+...+(\beta_1\alpha_{p-1}+\beta_p)x_p \\
&= (\beta_0+\beta_1\alpha_0)+\sum_{j=2}^{p}(\beta_1\alpha_{j-1}+\beta_j)x_j
\end{aligned} \tag{5.5}$$

$$\begin{aligned}
V(y|x_2,...,x_p) &= V[E(y|x_1,...,x_p)|x_2,...,x_p]+E[V(y|x_1,...,x_p)|x_2,...,x_p] \\
&= V(\beta_0+\beta_1x_1+...+\beta_px_p|x_2,..,x_p)+\sigma^2 \\
&= \beta_1^2\tau^2+\sigma^2.
\end{aligned} \tag{5.6}$$

By substituting Equation 5.3, Equation 5.4, Equation 5.5 and Equation 5.6 into Equation 5.2, the observed data likelihood is

$$\begin{aligned}
L = \prod_{i=1}^{n}\Bigg[&\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1x_{1i}-...-\beta_px_{pi})^2\right\} \\
&\frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{1}{2\tau^2}(x_{1i}-\alpha_0-\alpha_1x_{2i}-...-\alpha_{p-1}x_{pi})^2\right\}\Bigg]^{R_i}
\end{aligned}$$

$$\Big(\frac{1}{\sqrt{2\pi(\beta_1^2\tau^2+\sigma^2)}} \exp\Big[-\frac{1}{2(\beta_1^2\tau^2+\sigma^2)}\{y_i - (\beta_0+\beta_1\alpha_0) - \sum_{j=2}^{p}(\beta_1\alpha_{j-1}+\beta_j)x_{ji}\}^2\Big]\Big)^{1-R_i}. \tag{5.7}$$

Thus, the observed data log-likelihood is

$$\begin{aligned}
\log L &= \sum_{i=1}^{n} l_i(\beta,\alpha,\sigma,\tau) \\
&= \sum_{i=1}^{n} R_i\Big\{ -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{1i} - ... - \beta_p x_{pi})^2 \\
&\quad -\frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}(x_{1i} - \alpha_0 - \alpha_1 x_{2i} - ... - \alpha_{p-1}x_{pi})^2\Big\} \\
&\quad + (1-R_i)\Big\{ -\frac{1}{2}\log 2\pi(\beta_1^2\tau^2+\sigma^2) \\
&\quad -\frac{1}{2(\beta_1^2\tau^2+\sigma^2)}\left[y_i - (\beta_0+\beta_1\alpha_0) - \sum_{j=2}^{p}(\beta_1\alpha_{j-1}+\beta_j)x_{ji}\right]^2\Big\}.
\end{aligned} \tag{5.8}$$

To perform variable selection and estimation simultaneously, our goal is to optimize the SCAD type penalized likelihood function as follows.

$$\max_{\beta,\alpha,\sigma,\tau}\left\{ \sum_{i=1}^{n} l_i(\beta,\alpha,\sigma,\tau) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|) - n\sum_{j=1}^{p-1} p_\mu(|\alpha_j|)\right\}. \tag{5.9}$$

Because of the singularity property of SCAD function at the origin, the optimization function in Equation 5.9 is not differentiable with respect with β and α. Using the LLA described in 2.5, we can rewrite the optimization problem Equation 5.9 through the iterative process as follow,

$$\arg\max_{\beta,\alpha,\sigma,\tau}\left\{ \sum_{i=1}^{n} l_i(\beta,\alpha,\sigma,\tau) - n\sum_{j=1}^{p} p'_\lambda(|\beta_j^{(k)}|)|\beta_j| - n\sum_{j=1}^{p-1} p'_\mu(|\alpha_j^{(k)}|)|\alpha_j|\right\}, \tag{5.10}$$

where $\beta_j^{(k)}$ and $\alpha_j^{(k)}$ is the current step estimates.

Now the penalized likelihood function in Equation 5.11 becomes differentiable except for the origin. As demonstrated both empirically and theoretically in Zou and Li (2008), the one-step method is as efficient as the fully iterative method, provided that the initial estimators are reasonably good. If we set the initial estimate to be $(\beta^{(0)}, \alpha^{(0)}, \sigma^{(0)}, \tau^{(0)})$ the unpenalized maximum likelihood estimator and let $\theta = \beta, \alpha, \sigma, \tau$ be the vector for model parameters, the one-step MPLE can be obtained by

$$\theta^{(1)} = \arg\max_{\beta,\alpha,\sigma,\tau} \left\{ \sum_{i=1}^{n} l_i(\theta) - n \sum_{j=1}^{p} p_{\lambda}^{'}(|\beta_j^{(0)}|)|\beta_j| - \sum_{j=1}^{p-1} p_{\mu}^{'}(|\alpha_j^{(0)}|)|\alpha_j| \right\}. \tag{5.11}$$

Because the objective function on the right hand side of Equation 5.11 now has continuous second order derivatives except for the orgin, we can use the modified Newton-Raphson algorithm described in Fan and Li (2001) to solve the problem. Let

$$G(\beta, \alpha, \sigma, \tau) = \sum_{i=1}^{n} l_i(\beta, \alpha, \sigma, \tau) - n \sum_{j=1}^{p} p_{\lambda}^{'}(|\beta_j^{(0)}|)|\beta_j| - n \sum_{j=1}^{p-1} p_{\mu}^{'}(|\alpha_j^{(0)}|)|\alpha_j|.$$

The partial derivatives of G with respect to β's and α's are as follows.

$$\frac{\partial G}{\partial \beta_v} = \sum_{i=1}^{n} \left[\frac{\partial l_i(\beta, \alpha, \sigma, \tau)}{\partial \beta_v} - n p_{\lambda}^{'}(|\beta_v^{(0)}|) sgn(\beta_v) \right] \tag{5.12}$$

and

$$\frac{\partial G}{\partial \alpha_v} = \sum_{i=1}^{n} \left[\frac{\partial l_i(\beta, \alpha, \sigma, \tau)}{\partial \alpha_v} - n p_{\mu}^{'}(|\alpha_v^{(0)}|) sgn(\alpha_v) \right], \tag{5.13}$$

where $v = 1, \cdots, p$ in Equation 5.12 and $v = 1, \cdots, (p-1)$ in Equation 5.13.

In above two equations, we write out the first derivative of the log-likelihood in the following three equations.

$$\begin{aligned}\frac{\partial l_i}{\partial \beta_1} &= R_i\left\{\frac{x_{1i}}{\sigma^2}(y_i-\beta_0-\beta_1 x_{1i}-...-\beta_p x_{pi})\right\} \\ &\quad + (1-R_i)\left\{-\frac{\tau^2\beta_1}{\beta_1^2\tau^2+\sigma^2}-\frac{A}{4(\beta_1^2\tau^2+\sigma^2)^2}\right\}, \end{aligned} \tag{5.14}$$

where

$$\begin{aligned} A &= 2\left\{y_i-(\beta_0+\beta_1\alpha_0)-\sum_{j=2}^{p}(\beta_1\alpha_{j-1}+\beta_j)x_{ji}\right\}\cdot\left(-\alpha_0-\sum_{j=2}^{p}\alpha_{j-1}x_{ji}\right)\cdot 2(\beta_1^2\tau^2+\sigma^2) \\ &\quad -\left\{y_i-(\beta_0+\beta_1\alpha_0)-\sum_{j=2}^{p}(\beta_1\alpha_{j-1}+\beta_j)x_{ji}\right\}^2\cdot(4\tau^2\beta_1) \\ &= \left\{y_i-(\beta_0+\beta_1\alpha_0)-\sum_{j=2}^{p}(\beta_1\alpha_{j-1}+\beta_j)x_{ji}\right\}\cdot\left\{4(\beta_1^2\tau^2+\sigma^2)(-\alpha_0-\sum_{j=2}^{p}\right. \\ &\quad \left.\alpha_{j-1}x_{ji})-4\tau^2\beta_1\cdot\left[y_i-(\beta_0+\beta_1\alpha_0)-\sum_{j=2}^{p}(\beta_1\alpha_{j-1}+\beta_j)x_{ji}\right]\right\} \end{aligned} \tag{5.15}$$

$$\begin{aligned}\frac{\partial l_i}{\partial \beta_v} &= R_i\left\{\frac{x_{vi}}{\sigma^2}(y_i-\beta_0-\beta_1 x_{1i}-...-\beta_p x_{pi})\right\}+(1-R_i)\cdot \\ &\quad \left\{-\frac{x_{vi}}{\beta_1^2\tau^2+\sigma^2}\left[y_i-(\beta_0+\beta_1\alpha_0)-\sum_{j=2}^{p}(\beta_1\alpha_{j-1}+\beta_j)x_{ji}\right]\right\} \end{aligned} \tag{5.16}$$

$$\begin{aligned}\frac{\partial l_i}{\partial \alpha_v} &= R_i\left\{\frac{x_{(v+1)i}}{\tau^2}(x_{1i}-\alpha_0-\alpha_1 x_{2i}-...-\alpha_{p-1}x_{pi})\right\}+(1-R_i)\cdot \\ &\quad \left\{-\frac{[y_i-(\beta_0+\beta_1\alpha_0)-\sum_{j=2}^{p}(\beta_1\alpha_{j-1}+\beta_j)x_{ji}]\cdot(\beta_1 x_{(v+1)i})}{(\beta_1^2\tau^2+\sigma^2)}\right\}, \end{aligned} \tag{5.17}$$

where $v=2,...,p$ in Equation 5.16 and $v=1,...,(p-1)$ in Equation 5.17. We did not write out the other first order derivatives and the second order derivatives because they are tedious

and difficult to extend to more general cases with arbitrary missing patterns in covariates. We will use EM algorithm in the following section rather than directly maximize the observed data log-likelihood.

5.2 Variable Selection via Expectation-Maximization in A Simple Case

From the description in the previous section, directly solving Equation 5.11 is difficult and hard to be generalized. We apply the EM algorithm to solve the problem. The complete data likelihood for the problem described in section 5.1 is

$$\begin{aligned} L &= \prod_{i=1}^{n} L_i(\Theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{1i} - ... - \beta_p x_{pi})^2\right] \\ &\quad \cdot \frac{1}{\sqrt{2\pi}\tau} \cdot \exp\left[-\frac{1}{2\tau^2}(x_{1i} - \alpha_0 - \alpha_1 x_{2i} - ... - \alpha_{p-1} x_{pi})^2\right], \end{aligned} \tag{5.18}$$

where $\Theta = (\gamma, \eta)$ with $\gamma = (\sigma^2, \tau^2, \alpha_0, \beta_0)$ and $\eta = (\beta_1, ..., \beta_p, \alpha_1, ..., \alpha_{p-1})$.

In the EM algorithm, the E-step involves evaluating the Q-function defined as follows. The conditional expectation is taken with respect to the missing data given the observed data and the current estimated parameters, for which we denote $\gamma^{(0)}$ and $\eta^{(0)}$.

$$\begin{aligned} Q\left\{(\gamma, \eta)|(\gamma^{(0)}, \eta^{(0)})\right\} &= E\left[\log L|y, R(x_1), x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}\right] \\ &= \sum_{i=1}^{n} \left\{R_i \log L_i + (1 - R_i) E[\log L_i|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}]\right\} \end{aligned} \tag{5.19}$$

where, from Equation 5.18,

$$\begin{aligned} \log L_i &= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{1i} - ... - \beta_p x_{pi})^2 \\ &\quad - \frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}(x_{1i} - \alpha_0 - \alpha_1 x_{2i} - ... - \alpha_{p-1} x_{pi})^2 \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\tilde{y_i} - \beta_1 x_{1i})^2 \\
&\quad -\frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}(x_{1i} - \mu_i)^2, \qquad (5.20)
\end{aligned}
$$

with $\tilde{y_i} = y_i - \beta_0 - \beta_2 x_{2i} - ... - \beta_p x_{pi}$ and $\mu_i = \alpha_0 + \alpha_1 x_{2i} + ... + \alpha_{p-1} x_{pi}$.

Taking the conditional expectation with respect to the density of $(x_1|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)})$, we have

$$
\begin{aligned}
E\left[\log L_i|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}\right] &= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\log 2\pi\tau^2 \\
&\quad - \frac{1}{2\sigma^2}\Big\{\tilde{y_i}^2 - 2\tilde{y_i}\beta_1 E[x_1|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}] \\
&\quad + \beta_1^2 E[x_1^2|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}]\Big\} \\
&\quad - \frac{1}{2\tau^2}\Big\{\mu_i^2 - 2\mu_i E[x_1|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}] \\
&\quad + E[x_1^2|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}]\Big\} \qquad (5.21)
\end{aligned}
$$

The joint distribution $(y, x_1|x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)})$ is normal with mean

$$\mu_y^{(0)}(x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}) = (\beta_0^{(0)} + \beta_1^{(0)}\alpha_0^{(0)}) + \sum_{j=2}^{p}(\beta_1^{(0)}\alpha_{j-1}^{(0)} + \beta_j^{(0)})x_j,$$

$$\mu_{x_1}^{(0)}(x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}) = \alpha_0^{(0)} + \alpha_1^{(0)} x_{2i} + ... + \alpha_{p-1}^{(0)} x_{pi}$$

and variance matrix

$$\Sigma = \begin{bmatrix} \beta_1^{2(0)}\tau^{2(0)} + \sigma^{2(0)} & \sigma_{12} \\ \sigma_{12} & \tau^{2(0)} \end{bmatrix},$$

where

$$\sigma_{12} = Cov(y, x_1|x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)})$$

$$
\begin{aligned}
&= E(yx_1|x_2,...,x_p,\gamma^{(0)},\eta^{(0)}) - E(y|x_2,...,x_p,\gamma^{(0)},\eta^{(0)})\cdot E(x_1|x_2,...,x_p,\gamma^{(0)},\eta^{(0)}) \\
&= \beta_0^{(0)}(\alpha_0^{(0)}+\alpha_1^{(0)}x_{2i}+...+\alpha_{p-1}^{(0)}x_{pi}) \\
&\quad + \beta_1^{(0)}[\tau^{2(0)}+(\alpha_0^{(0)}+\alpha_1^{(0)}x_{2i}+...+\alpha_{p-1}^{(0)}x_{pi})^2] \\
&\quad + (\beta_2^{(0)}x_{2i}+...+\beta_p^{(0)}x_{pi})\cdot(\alpha_0^{(0)}+\alpha_1^{(0)}x_{2i}+...+\alpha_{p-1}^{(0)}x_{pi}) \\
&\quad - [\beta_0^{(0)}+\beta_1^{(0)}\alpha_0^{(0)}+\sum_{j=2}^{p}(\beta_1^{(0)}\alpha_{j-1}^{(0)}+\beta_j^{(0)})x_j]\cdot[\alpha_0^{(0)}+\alpha_1^{(0)}x_{2i}+...+\alpha_{p-1}^{(0)}x_{pi}] \\
&= \beta_1^{(0)}[\tau^{2(0)}+(\alpha_0^{(0)}+\alpha_1^{(0)}x_{2i}+...+\alpha_{p-1}^{(0)}x_{pi})^2] \\
&\quad + (\beta_2^{(0)}x_{2i}+...+\beta_p^{(0)}x_{pi})\cdot(\alpha_0^{(0)}+\alpha_1^{(0)}x_{2i}+...+\alpha_{p-1}^{(0)}x_{pi}) \\
&\quad - [\beta_1^{(0)}\alpha_0^{(0)}+\sum_{j=2}^{p}(\beta_1^{(0)}\alpha_{j-1}^{(0)}+\beta_j^{(0)})x_j]\cdot[\alpha_0^{(0)}+\alpha_1^{(0)}x_{2i}+...+\alpha_{p-1}^{(0)}x_{pi}] \\
&= \beta_1^{(0)}\tau^{2(0)}.
\end{aligned}
$$

Substitute σ_{12} with this quantity, we can write out the variance matrix as

$$
\Sigma = \begin{bmatrix} \beta_1^{2(0)}\tau^{2(0)}+\sigma^{2(0)} & \beta_1^{(0)}\tau^{2(0)} \\ \beta_1^{(0)}\tau^{2(0)} & \tau^{2(0)} \end{bmatrix}.
$$

The mean and variance of X_1 given $y, x_2, \ldots, x_p$ can be calculated as follows,

$$
\begin{aligned}
E[X_1|y,x_2,...,x_p,\gamma^{(0)},\eta^{(0)}] &= E[x_1|x_2,...,x_p,\gamma^{(0)},\eta^{(0)}] + cov(y,x_1|x_2,...,x_p,\gamma^{(0)},\eta^{(0)})\cdot \\
&\quad var^{-1}(y|x_2,...,x_p,\gamma^{(0)},\eta^{(0)})\cdot(y-E[y|x_2,...,x_p,\gamma^{(0)},\eta^{(0)}]) \\
&= (\alpha_0^{(0)}+\alpha_1^{(0)}x_{2i}+...+\alpha_{p-1}^{(0)}x_{pi})+\beta_1^{(0)}\tau^{2(0)}\cdot(\beta_1^{2(0)}\tau^{2(0)}+\sigma^{2(0)})^{-1} \\
&\quad \cdot[y-((\beta_0^{(0)}+\beta_1^{(0)}\alpha_0^{(0)})+\sum_{j=2}^{p}(\beta_1^{(0)}\alpha_{j-1}^{(0)}+\beta_j^{(0)})x_j)] \\
&= \mu_{x_1}^{(0)}+\frac{\beta_1^{(0)}\tau^{2(0)}}{\beta_1^{2(0)}\tau^{2(0)}+\sigma^{2(0)}}\cdot(y-\mu_y^{(0)}),
\end{aligned}
$$

and

$$\begin{aligned} V[X_1|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}] &= \tau^{2(0)} - \frac{(\beta_1^{(0)}\tau^{2(0)})^2}{\beta_1^{2(0)}\tau^{2(0)} + \sigma^{2(0)}} \\ &= \tau^{2(0)} \cdot \frac{\sigma^{2(0)}}{\beta_1^{2(0)}\tau^{2(0)} + \sigma^{2(0)}} \\ &\equiv v_0. \end{aligned}$$

The second order conditional expectation of X_1 is given by

$$\begin{aligned} E\left(x_1^2|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}\right) &= V\left[x_1|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}\right] + E^2\left(x_1|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}\right) \\ &= \tau^{2(0)} \cdot \frac{\sigma^{2(0)}}{\beta_1^{2(0)}\tau^{2(0)} + \sigma^{2(0)}} + \left[\mu_{x_1}^{(0)} + \frac{\beta_1^{(0)}\tau^{2(0)}}{\beta_1^{2(0)}\tau^{2(0)} + \sigma^{2(0)}} \cdot (y - \mu_y^{(0)})\right]^2. \end{aligned}$$

From above calculation, Equation 5.21 can be written as

$$\begin{aligned} & E\left(\log L_i|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}\right) \\ = & -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\log 2\pi\tau^2 \\ & - \frac{1}{2\sigma^2}\left\{\tilde{y_i}^2 - 2\tilde{y_i}\beta_1 E[x_1|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}] + \beta_1^2 E[x_1^2|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}]\right\} \\ & - \frac{1}{2\tau^2}\left\{\mu_i^2 - 2\mu_i E[x_1|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}] + E[x_1^2|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}]\right\} \\ = & -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left(\tilde{y_i}^2 - 2\tilde{y_i}\beta_1 \cdot c_1 + \beta_1^2 \cdot c_2\right) \\ & - \frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}\left(\mu_i^2 - 2\mu_i \cdot c_1 + c_2\right) \\ = & -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left\{(\tilde{y_i} - \beta_1 c_1)^2 + \beta_1^2(c_2 - c_1^2)\right\} \\ & - \frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}\left\{(\mu_i - c_1)^2 + (c_2 - c_1^2)\right\} \\ = & -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left\{(\tilde{y_i} - \beta_1 c_1)^2 + \beta_1^2 v_0\right\} \\ & - \frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}\left\{(\mu_i - c_1)^2 + v_0\right\} \end{aligned}$$

where c_1 and c_2 are constants, representing $E\left(x_1|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}\right)$ and $E\left(x_1^2|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}\right)$, respectively. Plug Equation 5.22 in Equation 5.19, the Q-function becomes

$$
\begin{aligned}
Q((\gamma,\eta)|(\gamma^{(0)},\eta^{(0)})) &= E[\log L|y, R(x_1), x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}] \\
&= \sum_{i=1}^{n}\{R_i \log L_i + (1-R_i)E[\log L_i|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}]\} \\
&= \sum_{i=1}^{n}\{R_i \cdot [-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\tilde{y_i} - \beta_1 x_{1i})^2 - \frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}(x_{1i} - \mu_i)^2] \\
&\quad + (1-R_i)\cdot E[\log L_i|y, x_2, ..., x_p, \gamma^{(0)}, \eta^{(0)}] \\
&= \sum_{i=1}^{n}\{R_i \cdot [-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\tilde{y_i} - \beta_1 x_{1i})^2 - \frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}(x_{1i} - \mu_i)^2] \\
&\quad + (1-R_i)\cdot\{-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\log 2\pi\tau^2 - \frac{1}{2\sigma^2}[(\tilde{y_i} - \beta_1 c_1)^2 + \beta_1^2 v_0] \\
&\quad - \frac{1}{2\tau^2}[(\mu_i - c_1)^2 + v_0]\}\} \\
&= \sum_{i=1}^{n}\{R_i \cdot [-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\log 2\pi\tau^2] - \frac{R_i}{2\sigma^2}(\tilde{y_i} - \beta_1 x_{1i})^2 - \frac{R_i}{2\tau^2}(x_{1i} - \mu_i)^2 \\
&\quad + (1-R_i)\cdot[-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\log 2\pi\tau^2] - \frac{1-R_i}{2\sigma^2}[(\tilde{y_i} - \beta_1 c_1)^2 + \beta_1^2 v_0] \\
&\quad - \frac{1-R_i}{2\tau^2}[(\mu_i - c_1)^2 + v_0]\} \\
&= \sum_{i=1}^{n}\Big\{R_i \cdot \left(-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\log 2\pi\tau^2\right) \\
&\quad - \frac{R_i}{2\sigma^2}\left(y_i - \beta_0 - \beta_2 x_{2i} - ... - \beta_p x_{pi} - \beta_1 x_{1i}\right)^2 \\
&\quad - \frac{R_i}{2\tau^2}\left(x_{1i} - \alpha_0 - \alpha_1 x_{2i} - ... - \alpha_{p-1} x_{pi}\right)^2 \\
&\quad + (1-R_i)\cdot\left(-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\log 2\pi\tau^2\right) \\
&\quad - \frac{1-R_i}{2\sigma^2}[(y_i - \beta_0 - \beta_2 x_{2i} - ... - \beta_p x_{pi} - \beta_1 c_1)^2 + \beta_1^2 v_0] \\
&\quad - \frac{1-R_i}{2\tau^2}[(\alpha_0 + \alpha_1 x_{2i} + ... + \alpha_{p-1} x_{pi} - c_1)^2 + v_0]\Big\}. \qquad (5.22)
\end{aligned}
$$

In M-step, we maximize $Q((\gamma,\eta)|(\gamma^{(0)},\eta^{(0)}))$ to obtain $(\gamma^{(1)},\eta^{(1)})$. Since γ contains the intercepts and variances of random error, they are estimated separately in linear regression

models. We first calculate the first and second derivatives of above objective function with respect to

$$\eta = (\beta_1, ..., \beta_p, \alpha_1, ..., \alpha_{p-1})\,.$$

The first and second partial derivatives of the Q-function defined in Equation 5.22 with respect to η can be written as follow.

$$\begin{aligned}\frac{\partial Q}{\partial \beta_1} &= \sum_{i=1}^{n}\Bigg\{\frac{R_i x_{1i}}{\sigma^2}(y_i - \beta_0 - \beta_2 x_{2i} - ... - \beta_p x_{pi} - \beta_1 x_{1i}) - \\ &\quad \frac{1-R_i}{\sigma^2}\left[(y_i - \beta_0 - \beta_2 x_{2i} - ... - \beta_p x_{pi} - \beta_1 c_1)(-c_1) + \beta_1(c_2 - c_1^2)\right]\Bigg\}\end{aligned}$$

$$\frac{\partial^2 Q}{\partial \beta_1^2} = \sum_{i=1}^{n}\left\{-\frac{R_i x_{1i}^2}{\sigma^2} - \frac{(1-R_i)c_2}{\sigma^2}\right\} \tag{5.23}$$

$$\begin{aligned}\frac{\partial Q}{\partial \beta_k} &= \sum_{i=1}^{n}\Bigg\{\frac{R_i x_{ki}}{\sigma^2}(y_i - \beta_0 - \beta_2 x_{2i} - ... - \beta_p x_{pi} - \beta_1 x_{1i}) - \\ &\quad \frac{1-R_i}{\sigma^2}(y_i - \beta_0 - \beta_2 x_{2i} - ... - \beta_p x_{pi} - \beta_1 c_1)(-x_{ki})\Bigg\}\end{aligned}$$

$$\frac{\partial^2 Q}{\partial \beta_k^2} = \sum_{i=1}^{n}\left\{-\frac{x_{ki}^2}{\sigma^2}\right\}, \tag{5.24}$$

where $k = 2, ..., p$.

$$\begin{aligned}\frac{\partial Q}{\partial \alpha_k} &= \sum_{i=1}^{n}\Bigg\{\frac{R_i x_{(k+1)i}}{\tau^2}(x_{1i} - \alpha_0 - \alpha_1 x_{2i} - ... - \alpha_{p-1} x_{pi}) - \\ &\quad \frac{(1-R_i)x_{(k+1)i}}{\tau^2}(x_{1i} + \alpha_0 + \alpha_1 x_{2i} + ... + \alpha_{p-1} x_{pi} - c_1)\Bigg\}\end{aligned}$$

$$\frac{\partial^2 Q}{\partial \alpha_k^2} = \sum_{i=1}^{n} \left\{ -\frac{x_{(k+1)i}^2}{\tau^2} \right\}, \tag{5.25}$$

where $k = 1, ..., (p-1)$. The remaining second mixed derivatives are as follow.

$$\frac{\partial^2 Q}{\partial \beta_1 \partial \beta_k} = \sum_{i=1}^{n} \left\{ \frac{-R_i x_{1i} x_{ki} - (1 - R_i) c_1 x_{ki}}{\sigma^2} \right\}, \tag{5.26}$$

where $k = 2, ..., p$;

$$\frac{\partial^2 Q}{\partial \beta_k \partial \beta_j} = \sum_{i=1}^{n} \left\{ \frac{-x_{ki} x_{ji}}{\sigma^2} \right\}, \tag{5.27}$$

where $k, j = 2, ..., p; k \neq j$;

$$\frac{\partial^2 Q}{\partial \alpha_k \partial \alpha_j} = \sum_{i=1}^{n} \left\{ \frac{-x_{(k+1)i} x_{(j+1)i}}{\tau^2} \right\}, \tag{5.28}$$

where $k, j = 1, ..., (p-1); k \neq j$ and

$$\frac{\partial^2 Q}{\partial \beta_j \partial \alpha_k} = 0, \tag{5.29}$$

where $j = 1, ..., p$ and $k = 1, ..., (p-1)$.

Compute $\eta^{(1)}$ using Newton-Raphson algorithm as follow

$$\eta^{(1)} = \eta^{(0)} - \frac{\dot{Q}_\eta((\eta, \gamma^{(0)})|\theta^{(0)})}{\ddot{Q}_{\eta\eta}((\eta, \gamma^{(0)}|\theta^{(0)})}. \tag{5.30}$$

Once we get $\eta^{(1)}$, we can update $\gamma^{(1)}$ as follows.

$$\beta_0^{(1)} = \bar{y} - \beta_1^{(1)} E(x_1|y, x_2, \cdots, x_p, \eta^{(1)}, \gamma^{(0)}) - \cdots - \beta_p^{(1)} \bar{x}_p, \tag{5.31}$$

$$\alpha_0^{(1)} = E(x_1|y, x_2, \cdots, x_p, \eta^{(1)}, \gamma^{(0)}) - \alpha_1^{(1)}\bar{x}_2 - \cdots - \alpha_{p-1}^{(1)}\bar{x}_p, \tag{5.32}$$

$$\sigma^{2(1)} = \frac{\sum_{i=1}^{n}(y_i - \beta_1^{(1)}E(x_{1i}|y, x_2, \cdots, x_p, \eta^{(1)}, \gamma^{(0)}) - \cdots - \beta_p^{(1)}x_{pi})^2}{n-1}, \tag{5.33}$$

and

$$\tau^{2(1)} = \frac{\sum_{i=1}^{n}(E(x_{1i}|y, x_2, \cdots, x_p, \eta^{(1)}, \gamma^{(0)}) - \alpha_1^{(1)}x_{2i} - \cdots - \alpha_{p-1}^{(1)}x_{pi})^2}{n-1}. \tag{5.34}$$

We iteratively repeat the process from Equation 5.30 to Equation 5.34 until the change from θ^k to θ^{k+1} is small enough. Let $\tilde{\theta}$ be the unpenalized estimator obtained by iteratively maximizing $Q(\theta|\theta^{(0)})$, where $\theta = (\beta, \alpha, \gamma)$. The variable selection problem is to maximize the following penalized pseudo likelihood function

$$Q(\theta|\tilde{\theta}) - n\sum_{j=1}^{2p-1} p_\lambda(|\eta_j|). \tag{5.35}$$

Following section 4.3, a Taylor series expansion at $\tilde{\theta}$ gives

$$\begin{aligned} n^{-1}Q(\theta|\tilde{\theta}) &\approx n^{-1}Q(\tilde{\theta}|\tilde{\theta}) + n^{-1}\dot{Q}_\theta(\tilde{\theta}|\tilde{\theta})(\theta - \tilde{\theta}) \\ &\quad + \frac{1}{2}(\theta - \tilde{\theta})^T\ddot{Q}(\tilde{\theta}|\tilde{\theta})(\theta - \tilde{\theta}) \\ &= n^{-1}Q(\tilde{\theta}|\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})^T\frac{\ddot{Q}(\tilde{\theta}|\tilde{\theta})}{n}(\theta - \tilde{\theta}). \end{aligned} \tag{5.36}$$

The last equation from Equation 5.36 is true because $Q'_\theta(\tilde{\theta}|\tilde{\theta}) = 0$. Therefore, the original optimization problem in Equation 5.35 is reformulated to maximize

$$(\theta - \tilde{\theta})^T\ddot{Q}(\tilde{\theta}|\tilde{\theta})(\theta - \tilde{\theta}) - n\sum_{j=1}^{2p-1} p_\lambda(|\eta_j|), \tag{5.37}$$

which is the same as

$$\left(\eta - \tilde{\eta}, \gamma - \tilde{\gamma} \right) \begin{pmatrix} \ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta}) & \ddot{Q}_{\eta\gamma}(\tilde{\theta}|\tilde{\theta}) \\ \ddot{Q}_{\gamma\eta}(\tilde{\theta}|\tilde{\theta}) & \ddot{Q}_{\gamma\gamma}(\tilde{\theta}|\tilde{\theta}) \end{pmatrix} \begin{pmatrix} \eta - \tilde{\eta} \\ \gamma - \tilde{\gamma} \end{pmatrix} - n \sum_{j=1}^{2p-1} p_\lambda(|\eta_j|), \tag{5.38}$$

Because our interest of parameter is η for variable selection, we try to maximize Equation 5.38 with fixed value of γ to get the one step maximum penalized pseudo likelihood estimator. In this case, maximize Equation 5.38 is then equivalent to maximize

$$\left(\eta - \tilde{\eta} + \ddot{Q}_{\eta\eta}^{-1}(\tilde{\theta}|\tilde{\theta})\ddot{Q}_{\eta\gamma}(\tilde{\theta}|\tilde{\theta})(\gamma - \tilde{\gamma})\right)^T \ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta}) \left(\eta - \tilde{\eta} + \ddot{Q}_{\eta\eta}^{-1}(\tilde{\theta}|\tilde{\theta})\ddot{Q}_{\eta\gamma}(\tilde{\theta}|\tilde{\theta})(\gamma - \tilde{\gamma})\right) - n \sum_{j=1}^{2p-1} p_\lambda(|\eta_j|), \tag{5.39}$$

with respect to η for a fixed γ. Notice that when we take the fixed value of γ to be the unpenalized maximum likelihood estimate $\tilde{\gamma}$, maximizing Equation 5.39 is equivalent to minimize

$$(\eta - \tilde{\eta})^T \frac{-\ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta})}{n} (\eta - \tilde{\eta}) + \sum_{j=1}^{2p-1} p_\lambda(|\eta_j|), \tag{5.40}$$

which is the penalized least square with penalty function $p_\lambda(\cdot)$ when $-\ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta})$ is positive-definite. Specifically, when $\ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta})$ is positive definite, it can be decomposed as $\ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta})$ as $\frac{-\ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta})}{n} = D^T D$ as below,

$$
\begin{aligned}
D^T D &= -\frac{1}{n}\begin{bmatrix}
\frac{\partial^2 Q}{\partial\beta_1^2} & \frac{\partial^2 Q}{\partial\beta_1\partial\beta_2} & \frac{\partial^2 Q}{\partial\beta_1\partial\beta_3} & \cdots & \frac{\partial^2 Q}{\partial\beta_1\partial\alpha_1} & \cdots & \frac{\partial^2 Q}{\partial\beta_1\partial\alpha_{p-1}} \\
\frac{\partial^2 Q}{\partial\beta_1\partial\beta_2} & \frac{\partial^2 Q}{\partial\beta_2^2} & \frac{\partial^2 Q}{\partial\beta_2\partial\beta_3} & \cdots & \frac{\partial^2 Q}{\partial\beta_2\partial\alpha_1} & \cdots & \frac{\partial^2 Q}{\partial\beta_2\partial\alpha_{p-1}} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{\partial^2 Q}{\partial\beta_p\partial\beta_1} & \frac{\partial^2 Q}{\partial\beta_p\partial\beta_2} & \frac{\partial^2 Q}{\partial\beta_p\partial\beta_3} & \cdots & \frac{\partial^2 Q}{\partial\beta_p\partial\alpha_1} & \cdots & \frac{\partial^2 Q}{\partial\beta_p\partial\alpha_{p-1}} \\
\frac{\partial^2 Q}{\partial\alpha_1\partial\beta_1} & \frac{\partial^2 Q}{\partial\alpha_1\partial\beta_2} & \frac{\partial^2 Q}{\partial\alpha_1\partial\beta_3} & \cdots & \frac{\partial^2 Q}{\partial\alpha_1^2} & \cdots & \frac{\partial^2 Q}{\partial\alpha_1\partial\alpha_{p-1}} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{\partial^2 Q}{\partial\alpha_{p-1}\partial\beta_1} & \frac{\partial^2 Q}{\partial\alpha_{p-1}\partial\beta_2} & \frac{\partial^2 Q}{\partial\alpha_{p-1}\partial\beta_3} & \cdots & \frac{\partial^2 Q}{\alpha_{p-1}\partial\alpha_1} & \cdots & \frac{\partial^2 Q}{\partial\alpha_{p-1}^2}
\end{bmatrix} \\
&= \begin{bmatrix}
\frac{\partial^2 Q}{\partial\beta_1^2} & \frac{\partial^2 Q}{\partial\beta_1\partial\beta_2} & \frac{\partial^2 Q}{\partial\beta_1\partial\beta_3} & \cdots & 0 & \cdots & 0 \\
\frac{\partial^2 Q}{\partial\beta_1\partial\beta_2} & \frac{\partial^2 Q}{\partial\beta_2^2} & \frac{\partial^2 Q}{\partial\beta_2\partial\beta_3} & \cdots & 0 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{\partial^2 Q}{\partial\beta_p\partial\beta_1} & \frac{\partial^2 Q}{\partial\beta_p\partial\beta_2} & \frac{\partial^2 Q}{\partial\beta_p\partial\beta_3} & \cdots & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & \frac{\partial^2 Q}{\partial\alpha_1^2} & \cdots & \frac{\partial^2 Q}{\partial\alpha_1\partial\alpha_{p-1}} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & \cdots & \frac{\partial^2 Q}{\alpha_{p-1}\partial\alpha_1} & \cdots & \frac{\partial^2 Q}{\partial\alpha_{p-1}^2}
\end{bmatrix}.
\end{aligned}
$$

Therefore, Equation 5.37 can be reformulated as follows.

$$
\begin{aligned}
(43) &= \arg\min_\eta \left\{ (\eta-\tilde{\eta})^T \frac{-\ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta})}{n}(\eta-\tilde{\eta}) + \sum_{j=1}^{2p-1} p_\lambda(|\eta_j|) \right\} \\
&= \arg\min_\eta \left\{ (\eta-\tilde{\eta})^T D^T D(\eta-\tilde{\eta}) + \sum_{j=1}^{2p-1} p_\lambda(|\eta_j|) \right\} \\
&= \arg\min_\eta \left\{ (D\eta-D\tilde{\eta})^T (D\eta-D\tilde{\eta}) + \sum_{j=1}^{2p-1} p_\lambda(|\eta_j|) \right\}
\end{aligned}
$$

$$= \arg\min_{\eta} \left\{ (y^* - x^*\eta)^T (y^* - x^*\eta) + \sum_{j=1}^{2p-1} p_\lambda(|\eta_j|) \right\}, \tag{5.41}$$

where $y^* = -D\tilde{\eta}$ and $x^* = -D$. Therefore, it is the same as solving a variable selection problem with SCAD penalty under the linear regression model with fully observed pseudo-data.

For the algorithm described in 4.4, since it does not involve second order derivatives of the Q-function, the one step penalized maximum likelihood estimator can be computed directly on η by minimizing

$$\frac{1}{2}(\eta - \tilde{\eta})^T \left\{ \sum_{i=1}^{n} \dot{Q}_{i(\eta)}^T(\tilde{\theta} \mid \tilde{\theta}) \dot{Q}_{i(\eta)}(\tilde{\theta} \mid \tilde{\theta}) \right\} (\eta - \tilde{\eta}) + n \sum_{j=1}^{2p-1} p_{j\lambda}(|\eta_j|), \tag{5.42}$$

5.3 Expectation Maximization Algorithm for Multivariate Normal

In this section, we combine the outcome and covariates into a single vector for the illustration of EM algorithm for multivariate normal. Let $y = (x_1, \ldots, x_p)^T \sim N(\mu, \Sigma)$ and $y_1, \ldots, y_n \sim^{i.i.d} N(\mu, \Sigma)$. The full likelihood of the problem is

$$\begin{aligned} L &= \prod_{i=1}^{n} L_i(\Theta) \\ &= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{\frac{1}{2}}} \cdot exp\left\{ -\frac{1}{2}(y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right\}, \end{aligned} \tag{5.43}$$

where $\Theta = (\mu, \Sigma)$. The log-likelihood is

$$\begin{aligned} l(\mu, \Sigma) &= logL \\ &= -\frac{pn}{2} \log 2\pi - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \\ &= -\frac{pn}{2} \log 2\pi - \frac{n}{2} \log |\Sigma| - \frac{1}{2} tr \left\{ \Sigma^{-1} \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)^T \right\}. \end{aligned} \tag{5.44}$$

In E-step of EM algorithm, the conditional expectation of the complete-data likelihood is

$$\begin{aligned} Q(\mu, \Sigma|\mu^{(0)}, \Sigma^{(0)}) &= -\frac{pn}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| \\ &\quad - \frac{1}{2}tr\left\{\Sigma^{-1}\sum_{i=1}^{n} E[(y_i-\mu)(y_i-\mu)^T]\Big|y_i^{obs}, \mu^{(0)}, \Sigma^{(0)}\right\}, \end{aligned} \tag{5.45}$$

where y_i^{obs} is observed part of the vector $y_i = (x_{i1}, \ldots, x_{ip})$.

Thus, to estimate $Q(\mu, \Sigma|\mu^{(0)}, \Sigma^{(0)})$, we first need to get the following two items.

1. $E(x_{ij}^2|y_i^{obs}, \mu^{(0)}, \Sigma^{(0)})$
2. $E(x_{ij}x_{ik}|y_i^{obs}, \mu^{(0)}, \Sigma^{(0)})$

Suppose that y is a $p \times 1$ random vector distributed as $N(\mu, \Sigma)$. We partition as $y^T = (z_1^T, z_2^T)$, where z_1 and z_2 are sub-vectors of lengths p_1 and $p_2 = p - p_1$, respectively. It is well known that their marginal distributions are partitions of the μ and Σ, i.e., $\mu^T = (\mu_1^T, \mu_2^T)$ and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Further, the conditional distributions are also normal; for example, the distribution of z_2 given z_1 in normal with mean

$$\begin{aligned} E(z_2|z_1) &= (\alpha_{2.1}) + B_{2.1} * z_1 \\ &= (\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1) + (\Sigma_{21}\Sigma_{11}^{-1}) * z_1 \end{aligned}$$

and covariance

$$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

By Sweep Operator, we can conveniently calculate the above values. We arrange the parameters $\theta = (\mu, \Sigma)$ as a $(p+1) \times (p+1)$ matrix in the following manner,

$$\begin{aligned} \theta &= \begin{bmatrix} 1 & \mu \\ \mu & \Sigma \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mu_1^T & \mu_2^T \\ \mu_1 & \Sigma_{11} & \Sigma_{12} \\ \mu_2 & \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \end{aligned}$$

If we sweep this θ matrix on positions $1, 2, \ldots, p_1$; the resulting matrix is

$$\begin{aligned} SWP[1, \ldots, p_1]\theta &= \begin{bmatrix} 1 - \mu_1^T \Sigma_{11}^{-1} \mu_1 & \mu_1^T \Sigma_{11}^{-1} & \mu_2^T - \mu_1^T \Sigma_{11}^{-1} \Sigma_{12} \\ \Sigma_{11}^{-1} \mu_1 & -\Sigma_{11}^{-1} & \Sigma_{11}^{-1} \Sigma_{12} \\ \mu_2 - \Sigma_{21} \Sigma_{11}^{-1} \mu_1 & \Sigma_{21} \Sigma_{11}^{-1} & \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \end{bmatrix} \\ &= \begin{bmatrix} 1 - \mu_1^T \Sigma_{11}^{-1} \mu_1 & \mu_1^T \Sigma_{11}^{-1} & \alpha_{2.1}^T \\ \Sigma_{11}^{-1} \mu_1 & -\Sigma_{11}^{-1} & B_{2.1}^T \\ \alpha_{2.1} & B_{2.1} & \Sigma_{22.1} \end{bmatrix}. \end{aligned}$$

It is easy to see that the information used to calculate conditional expectation is in sweeped columns and unsweeped rows; and the covariance is in unsweeped column and rows. For our multivariate normal problem with missing data, if row i is in missingness pattern s, let $O(s)$ and $M(s)$ denote observed and missing columns, respectively. And denote A the swept parameter matrix

$$A = SWP[O(s)]\theta,$$

and a_{jk} the (j,k)th element of A, $j,k = 0,1,\ldots,p$. The first two moments of y_i^{mis} with respect to $P(y^{mis}|y^{obs},\theta)$ are given by:

$$E(x_{ij}|y_i^{obs},\mu^{(0)},\Sigma^{(0)}) = a_{0j} + \sum_{k\in O(s)} a_{kj}x_{ik}$$
$$cov(x_{ij}x_{ik}|y_i^{obs},\mu^{(0)},\Sigma^{(0)}) = a_{jk}$$

for each each $j,k \in M(s)$. For any $j \in O(s)$, the moments are

$$E(x_{ij}|y_i^{obs},\mu^{(0)},\Sigma^{(0)}) = x_{ij}$$
$$cov(x_{ij}x_{ik}|y_i^{obs},\mu^{(0)},\Sigma^{(0)}) = 0.$$

By applying the relation

$$\begin{aligned} E(x_{ij}x_{ik}|y_i^{obs},\mu^{(0)},\Sigma^{(0)}) &= cov(x_{ij}x_{ik}|y_i^{obs},\mu^{(0)},\Sigma^{(0)}) + \\ &\quad E(x_{ij}|y_i^{obs},\mu^{(0)},\Sigma^{(0)})E(x_{ik}|y_i^{obs},\mu^{(0)},\Sigma^{(0)}) \end{aligned}$$

it follows that

$$E(x_{ij}|y_i^{obs},\mu^{(0)},\Sigma^{(0)}) = \begin{cases} x_{ij} & \text{for } j \in O(s), \\ x_{ij}^* & \text{for } j \in M(s). \end{cases}$$

and

$$E(x_{ij}x_{ik}|y_i^{obs},\mu^{(0)},\Sigma^{(0)}) = \begin{cases} x_{ij}x_{ik} & \text{for } j,k \in O(s), \\ x_{ij}^*x_{ik} & \text{for } j \in M(s), k \in O(s), \\ a_{jk} + x_{ij}^*x_{ik}^* & \text{for } j,k \in M(s). \end{cases}$$

where

$$x^*_{ij} = a_{0j} + \sum_{k \in O(s)} a_{kj} x_{ik}.$$

5.4 Alternative Parameterization in Variable Selection

We can write the multivariate normal model in the form of p successive linear regression models as follow.

$x_1|x_2, \ldots, x_p \sim N(\eta_1, \sigma_1^2)$

$x_2|x_3, \ldots, x_p \sim N(\eta_2, \sigma_2^2)$

$\ldots$

$x_{p-1}|x_p \sim N(\eta_{p-1}, \sigma_{p-1}^2)$

$x_p \sim N(\eta_p, \sigma_p^2)$,

where $\eta_j = \beta_{j0} + \beta_{j,(j+1)} x_{j+1} + \ldots + \beta_{j,p} x_p$, $j = 1, \ldots, (p-1)$.

The log-likelihood can be rewritten as a function of (β, σ) as follows.

$$\begin{aligned}
l(\beta, \sigma) &= \sum_{i=1}^{n} \{ -\frac{1}{2} log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_{i1} - \eta_1)^2 \\
&\quad - \frac{1}{2} log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_{i2} - \eta_2)^2 \\
&\quad \ldots\ldots \\
&\quad - \frac{1}{2} log(2\pi\sigma_{p-1}^2) - \frac{1}{2\sigma_{p-1}^2}(x_{i(p-1)} - \eta_{p-1})^2 \\
&\quad - \frac{1}{2} log(2\pi\sigma_p^2) - \frac{1}{2\sigma_p^2}(x_{ip} - \eta_p)^2 \} \\
&= \sum_{i=1}^{n} \{ -\frac{1}{2} log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_{i1} - (\beta_{10} + \beta_{12} x_{i2} + \ldots + \beta_{1p} x_{ip}))^2 \\
&\quad - \frac{1}{2} log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_{i2} - (\beta_{20} + \beta_{23} x_{i3} + \ldots + \beta_{2p} x_{ip}))^2 \\
&\quad \ldots\ldots
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{2} log(2\pi\sigma_{p-1}^2) - \frac{1}{2\sigma_{p-1}^2}(x_{i(p-1)} - (\beta_{(p-1)0} + \beta_{(p-1)p}x_{pi}))^2 \\
&- \frac{1}{2} log(2\pi\sigma_p^2) - \frac{1}{2\sigma_p^2}(x_{ip} - \mu_p)^2\}, \qquad (5.46)
\end{aligned}
$$

where $\beta = (\beta_{jk}, j = 1, \cdots, p-1; k = j, \cdots, p)$ and $\sigma = (\sigma_1, \cdots, \sigma_p)$.

The Q-function is the expected full data log-likelihood conditional on the observed data under the currently estimated model as follows.

$$
\begin{aligned}
Q(\theta|(x^{obs}, \theta^{(0)})) \quad = \quad & \sum_{i=1}^{n}\{-\frac{1}{2} log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}E[(x_{i1} - (\beta_{10} + \beta_{12}x_{i2} + \ldots + \beta_{1p}x_{ip}))^2|(x^{obs}, \theta^{(0)})] \\
& - \frac{1}{2} log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}E[(x_{i2} - (\beta_{20} + \beta_{23}x_{i3} + \ldots + \beta_{2p}x_{ip}))^2|(x^{obs}, \theta^{(0)})] \\
& \ldots\ldots \\
& - \frac{1}{2} log(2\pi\sigma_{p-1}^2) - \frac{1}{2\sigma_{p-1}^2}E[(x_{i(p-1)} - (\beta_{(p-1)0} + \beta_{(p-1)p}x_{ip}))^2|(x^{obs}, \theta^{(0)})] \\
& - \frac{1}{2} log(2\pi\sigma_p^2) - \frac{1}{2\sigma_p^2}E[(x_{ip} - \mu_p)^2|(x^{obs}, \theta^{(0)})]\}, \qquad (5.47)
\end{aligned}
$$

where $\theta = (\eta, \gamma)$ and $\eta = (\beta_{jk}, j = 1, \ldots, p-1; k = j, \ldots, p)$, $\gamma = (\sigma_j, j = 1, \ldots, p)$. The first and second derivatives of Q with respect to regression model coefficients are as follow,

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_{jk}} \quad = \quad & -\frac{1}{\sigma_j^2}\sum_{i=1}^{n}\Big\{\beta_{jk}E[x_{ik}^2|(x^{obs}, \theta^{(0)})] \\
& - E[x_{ik}(x_{ij} - \beta_{j0} - \beta_{j2}x_{i2} - \ldots - \beta_{j(k-1)}x_{i(k-1)} - \beta_{j(k+1)}x_{i(k+1)} \\
& - \ldots - \beta_{jp}x_{ip})|(x^{obs}, \theta^{(0)})]\Big\} \qquad (5.48)
\end{aligned}
$$

The second derivatives appear as

$$
\frac{\partial^2 Q}{\partial \beta_{jk}^2} \quad = \quad -\frac{1}{\sigma_j^2}\sum_{i=1}^{n}\left\{E\left[x_{ik}^2\middle|(x^{obs}, \theta^{(0)})\right]\right\}, \qquad (5.49)
$$

and

$$\frac{\partial^2 Q}{\partial\beta_{jk}\partial\beta_{jl}} = -\frac{1}{\sigma_j^2}\sum_{i=1}^{n}\left\{E\left[x_{ik}x_{il}\middle|(x^{obs},\theta^{(0)})\right]\right\}, \tag{5.50}$$

where $j = 1 \sim (p-1); l, k = j \sim p; l \neq k$ and

$$\frac{\partial^2 Q}{\partial\beta_{jk}\partial\beta_{j'l}} = 0 \tag{5.51}$$

when $j \neq j'$. Then, $\eta^{(1)}$ can be computed by the following Newton-Raphson iteration

$$\eta^{(1)} = \eta^{(0)} - \frac{\dot{Q}_\eta((\eta,\gamma^{(0)})|\theta^{(0)})}{\ddot{Q}_{\eta\eta}((\eta,\gamma^{(0)}|\theta^{(0)})}. \tag{5.52}$$

$\gamma^{(1)}$ then can be updated similarly as in Equation 5.30 to Equation 5.34 for each regression models described in Equation 5.46. Once the unpenalized maximum likelihood estimator $\hat{\theta}$ is obtained, with fixed $\hat{\gamma}$ values, the one step penalized maximum likelihood estimator can be computed by minimizing

$$\frac{1}{2}(\eta-\hat{\eta})^T\left\{\sum_{i=1}^{n}\dot{Q}_{i(\eta)}^T(\hat{\theta}\mid\hat{\theta})\dot{Q}_{i(\eta)}(\hat{\theta}\mid\hat{\theta})\right\}(\eta-\hat{\eta}) + n\sum_{j=1}^{K}p_{j\lambda}(|\eta_j|), \tag{5.53}$$

where $K = \frac{p(p-1)}{2}$ is the number of coefficients in $p-1$ regression models.

5.5 A Simulation Study

In this section, we conduct a simulation study to demonstrate the performance of our algorithm. In the simulation we considered a multivariate normal data with a sparse inverse covariance structure with different sample size. Since we pre-specified the covariance structure, the consecutive regression coefficients are determined as well so that we can evaluate the performance of our algorithm to see if it can capture the true model. We used the

LLA algorithm of (56) with LARS algorithm to compute the maximum penalized likelihood estimates with SCAD penalty as in the objective function Equation 5.41. For each iteration, we select the tuning parameter by BIC criteria, described in 4.5.

Model 1a (Multivariate Normal) We let $\mathbf{x} \sim N(0, \Sigma)$ be a $n \times p$ matrix. We will simulate three different sample size $n = 200, 400$ and 800. Let $p = 10$. The covariance matrix $\Sigma_{10\times 10}$ is constructed as follows. Set A be a correlation matrix with pairwise correlation $0.5^{|j_1-j_2|}$ between x_{j1} and x_{j2} for $j_1, j_2 \leq 5$; otherwise set to 0. The choice is to make sure the true coefficients are sparse. Let $\Sigma = A^T A$, to guarantee the positive definite of covariance matrix. Since we have 10 variables here, the number of consecutive regression coefficients except intercept is $\frac{p(p-1)}{2} = 45$. Once the mean and covariance structure is fixed for a multivariate normal data, the true value of coefficients in the regression models are known to us. Under the covariance structure just described, the number of non-zero coefficients is 10. They are $\beta_{1,2}, \beta_{1,3}, \beta_{1,4}, \beta_{1,5}, \beta_{2,3}, \beta_{2,4}, \beta_{2,5}, \beta_{3,4}, \beta_{3,5}$ and $\beta_{4,5}$. Their true values and estimates are listed in Table V. To impose missing at random mechanism into simulation, the following selection procedure is assumed. Covariates $x_5 \sim x_{10}$ are set to be fully observed, for x_1 to x_4, the probability of an observed value in x_j is

$$p(R_{x_j} = 1) = \frac{exp(\alpha_j + \alpha^T \mathbf{x_{5\sim 10}}))}{1 + exp(\alpha_j + \alpha^T \mathbf{x_{5\sim 10}}))},$$

where $\mathbf{x_{5\sim 10}} = (x_5, x_6, x_7, x_8, x_9, x_{10})^T$, $\alpha_1 = 1.7$, $\alpha_2 = 1.6$, $\alpha_3 = 1.4$, $\alpha_4 = 1.5$ and $\alpha = (0.8, -0.3, 0.5, -0.05, 0.8, 0.6)^T$. The average missing proportions for $x_j, j = 1 \sim 4$ are 20%, 24%, 28% and 26%, respectively. We repeated the simulation for 1000 times for both methods described in section 4.3 (Q-method) and section 4.4 (L-method).

TABLE IV

SIMULATION RESULT FOR MULTIVARITE NORMAL DATA (MAR)

			No. of Zeros/Non-zeros		Proportion of		
Method	Sample Size	MRME	C (sd)	IC (sd)	Under-fit	Correct-fit	Over-fit
	n=200	0.59	9.60 (0.64)	0.04 (0.22)	0.33	0.64	0.04
Q-func	n=400	0.65	9.67 (0.61)	0.00 (0.00)	0.25	0.75	0.00
	n=800	0.73	9.76 (0.51)	0.00 (0.00)	0.21	0.79	0.00
	n=200	0.53	9.71 (0.63)	0.11 (0.34)	0.21	0.69	0.01
L-func	n=400	0.56	9.75 (0.59)	0.01 (0.10)	0.19	0.80	0.01
	n=800	0.65	9.82 (0.52)	0.00 (0.00)	0.13	0.87	0.00

For linear models, model error for $\hat{\mu} = x^T\beta$ is $(\hat{\beta} - \beta)^T E(x^T x)(\hat{\beta} - \beta)$. Simulation results are summarized in Table IV, in which MRME stands for median of ratio of model error of a selected model to that of the unpenalized estimates under the full model. Columns "C" and "IC" measures models' complexity. Column "C" calculates the average number of non-zero coefficients correctly estimated to be non-zeros and column "IC" calculates the average number of zero coefficients incorrectly estimated to be non-zero. In an ideal case, "C" should equal to 10 (identify all non-zero coefficients) and "IC" should equal to 0. In the column labeled "Under-fit", we presented the proportion of excluding any non-zero efficients in one thousand replications. Similarly, we reported the probability of selecting the exact subset model and the probability of including all 10 significant variables and some noise variables in the columns "Correct-fit" and "Over-fit", respectively.

As it can be seen from Table IV, the sparse estimates from both methods dramatically reduce model error and have a greater chance to identify the true model as sample size increases. When sample size is small ($n = 200$), the method described in section 4.3 using second derivatives of the Q-function (Q-func) in the objective function with BIC selection criteria

TABLE V

PART I: MEAN OF REGRESSION COEFFICIENTS FOR PENALIZED METHOD USING THE Q-FUNCTION

Coefficient	n=200	n=400	n=800	true value
β_1	1.52 (0.14)	1.53 (0.12)	1.56 (0.09)	1.45
β_2	-1.48 (0.31)	-1.49 (0.27)	-1.54 (0.21)	-1.47
β_3	1.10 (0.39)	1.11 (0.35)	1.19 (0.27)	1.16
β_4	-0.54 (0.27)	-0.55 (0.23)	-0.59 (0.16)	-0.64
β_5	1.20 (0.20)	1.21 (0.19)	1.24 (0.17)	1.33
β_6	-0.91 (0.39)	-0.93 (0.37)	-1.00 (0.33)	-1.17
β_7	0.40 (0.33)	0.42 (0.31)	0.47 (0.27)	0.67
β_8	1.09 (0.17)	1.11 (0.13)	1.13 (0.08)	1.14
β_9	-0.58 (0.23)	-0.61 (0.18)	-0.64 (0.11)	-0.71
β_{10}	0.73 (0.04)	0.73 (0.03)	0.73 (0.02)	0.80

on average under-fit the model by excluding $0.4(= 10 - 9.6)$ non-zero variables from the final model. This inaccuracy is improved with the method in section 4.4, which uses first derivatives of Q-function to approximate the second derivative of the observed data log-likelihood. When sample size doubles to 400, the proposed algorithm on average identify 11% more in the correct fit, in comparison, Q-method increases 9%. When sample size increases to n=800, chance of a correct fit increases from 80% to 87% and more accurate coefficient estimates are obtained. See Table V for the details. For the Q-Method, the rate of correct fit increases from 75% to 79%.

Model 1b (Simulation of Adaptation to Data with Binary Variable) Using the same mean and covariance structure to generate a multivariate normal data as in *Model 1a*: that is $\mathbf{x} \sim N(0, \Sigma)$ is a $n \times p$ matrix, where $n = 400, 800$ and $p = 10$. Adding a binary variable y into x so that the dimension of new dataset is n by $p + 1$. Then, the number of consecutive regression coefficients is $\frac{(p+1)p)}{2} = 55$. We simulated three scenarios of dependence. In the first

TABLE VI

PART II: MEAN OF REGRESSION COEFFICIENTS FOR PENALIZED METHOD USING THE L-FUNCTION

Coefficient	n=200	n=400	n=800	true value
β_1	1.59 (0.10)	1.59 (0.06)	1.59 (0.05)	1.45
β_2	-1.62 (0.21)	-1.60 (0.15)	-1.60 (0.12)	-1.47
β_3	1.26 (0.28)	1.23 (0.37)	1.24 (0.21)	1.16
β_4	-0.61 (0.26)	-0.58 (0.26)	-0.59 (0.22)	-0.64
β_5	1.36 (0.34)	1.36 (0.27)	1.36 (0.24)	1.33
β_6	-1.21 (0.49)	-1.21 (0.46)	-1.19 (0.44)	-1.17
β_7	0.63 (0.48)	0.64 (0.50)	0.62 (0.48)	0.67
β_8	1.11 (0.36)	1.12 (0.27)	1.13 (0.24)	1.14
β_9	-0.62 (0.40)	-0.63 (0.34)	0.62 (0.32)	-0.71
β_{10}	0.72 (0.25)	0.72 (0.19)	0.73 (0.16)	0.80

case, y is distributed as *Bernoulli* and depends on none of the covariates. In the second case, y depends on 1 covariate in $\mathbf{x}$ as $y|\mathbf{x} \sim Bernoulli\{p(2x_3)\}$ where $p(u) = \frac{e^u}{1+e^u}$. In the third case, y depends on x_1 and the model is $y|\mathbf{x} \sim Bernoulli\{p(2x_1)\}$. Missing mechanism was simulated in the same way as in *model 1a*. Simulation results are shown in Table Table VII. The column "Prop-Log" denotes the proportion of correctly identification of regression model for the outcome variable.

From Table VII, we can see our algorithm for variable selection using a multivariate normal does not perform well when the assumption of multivariate normal is incorrect. In the first scenario when the binary outcome variable is independent with other covariates, our simulation shows the proposed algorithm can still identify the true model with more than 80%. In the second case when the outcome variable depends on x_3 with a large coefficient, the algorithm yields a biased estimate for the outcome regression model but identify more than 70% of the true model. In the last case when the outcome depends on the covariates subject

TABLE VII

RESULTS FOR SIMULATION OF ADAPTATION TO DATA WITH BINARY VARIABLE

			No. of Zeros/Non-zeros		Proportion of		
Size	Model	Prop-Log	C (sd)	IC (sd)	Under-fit	Correct-fit	Over-fit
	Case1	0.95	9.75 (0.62)	0.14 (0.62)	0.16	0.79	0.05
n=400	Case2	0.72	10.67 (0.61)	0.12 (0.41)	0.26	0.70	0.03
	Case3	0.27	10.83 (0.88)	0.13 (0.39)	0.68	0.27	0.05
	Case1	0.99	9.78 (0.55)	0.01 (0.13)	0.16	0.83	0.01
n=800	Case2	0.76	10.71 (0.55)	0.04 (0.24)	0.24	0.75	0.01
	Case3	0.39	11.03 (0.87)	0.00 (0.00)	0.61	0.39	0.00

to missing values, the proposed algorithm for multivariate normal data fail to identify the correct logistic regression model, though it can identify most non-zero coefficients in the linear regression models for continuous variables. The simulation shows that though the algorithm under multivariate normal model assumption appears for some cases to be able to identify the models for the logistic regression model, the estimates appears to be biased. The performance of the misspecified model becomes much worse when missing covariates are involved.

CHAPTER 6

VARIABLE SELECTION FOR NON-LINEAR REGRESSION WITH MISSING COVARIATES

6.1 Logistic Regression with Missing Continuous Covariates

Suppose we have a binary outcome $\mathbf{y}$ and covariates $\mathbf{x_1}, \ldots, \mathbf{x_n} \sim N(\mu, \Sigma)$. Using the consecutive conditional model of section 4.3, the full data likelihood is

$$\begin{aligned} l(\beta, \sigma) &= \sum_{i=1}^{n} \{ y_i \log p_i + (1 - y_i) \log(1 - p_i) \\ &\quad - \frac{1}{2} \log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_{i1} - \eta_{i1})^2 \\ &\quad - \frac{1}{2} \log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_{i2} - \eta_{i2})^2 \\ &\quad \ldots\ldots \\ &\quad - \frac{1}{2} \log(2\pi\sigma_{p-1}^2) - \frac{1}{2\sigma_{p-1}^2}(x_{i(p-1)} - \eta_{i(p-1)})^2 \\ &\quad - \frac{1}{2} \log(2\pi\sigma_p^2) - \frac{1}{2\sigma_p^2}(x_{ip} - \eta_{ip})^2 \}, \end{aligned} \tag{6.1}$$

where $p_i = \frac{\exp(\mathbf{x}_i\beta_y)}{1+\exp(\mathbf{x}_i\beta_y)}$, $\mathbf{x}_i\beta_y = \beta_{y0} + \beta_{y1}x_{i1} + \ldots + \beta_{yp}x_{ip}$ and $\eta_{ij} = \beta_{j0} + \beta_{j,(j+1)}x_{i(j+1)} + \ldots + \beta_{j,p}x_{ip}$, $j = 1, \ldots, p-1$.

6.1.1 Logistic Regression with One Missing Covariate

Consider first the simple case where only x_1 is subject to missing. Let $\theta = (\gamma, \eta)$, where $\eta = (\beta_{jk}, j = 1, \ldots, p-1; k = j, \ldots, p)$ and $\gamma = (\sigma_{j'}^2, \beta_{y0}, \beta_{j'0}, j' = 1, \ldots, p-1)$. Then

the expected full data log-likelihood conditional on the observed data and current parameter estimates $\theta^{(0)}$ is

$$\begin{aligned} Q\left\{\theta|(y,x^{obs},\theta^{(0)})\right\} &= \sum_{i=1}^{n}\Big[y_i\{\beta_{y0}+\beta_{y1}E(x_{i1}|y_i,x_i^{obs},\theta^{(0)})+\ldots+\beta_{yp}x_{ip}\} \\ &\quad -E\{\log(1+e^{\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip}})|y_i,x_i^{obs},\theta^{(0)}\} \\ &\quad -\frac{1}{2}\log(2\pi\sigma_1^2)-\frac{1}{2\sigma_1^2}E\{(x_{i1}-(\beta_{10}+\beta_{12}x_{i2}+\ldots+\beta_{1p}x_{ip}))^2|y_i,x_i^{obs},\theta^{(0)}\} \\ &\quad -\frac{1}{2}\log(2\pi\sigma_2^2)-\frac{1}{2\sigma_2^2}E\{(x_{i2}-(\beta_{20}+\beta_{23}x_{i3}+\ldots+\beta_{2p}x_{ip}))^2|y_i,x_i^{obs},\theta^{(0)}\} \\ &\quad \ldots\ldots \\ &\quad -\frac{1}{2}\log(2\pi\sigma_{p-1}^2)-\frac{1}{2\sigma_{p-1}^2}E\{(x_{i(p-1)}-(\beta_{(p-1)0}+\beta_{(p-1)p}x_{ip}))^2|y_i,x_i^{obs},\theta^{(0)}\} \\ &\quad -\frac{1}{2}\log(2\pi\sigma_p^2)-\frac{1}{2\sigma_p^2}E[(x_{ip}-\mu_p)^2|y_i,x_i^{obs},\theta^{(0)}\Big], \end{aligned} \tag{6.2}$$

where $x_i^{obs}=(x_{i2},\ldots,x_{ip})$ if x_{i1} is missing and $x_i^{obs}=(x_{i1},\ldots,x_{ip})$ if x_{i1} is observed.

The first derivatives of above expected full-data log-likelihood conditional on the observed data are

1. For β_{yl},

$$\begin{aligned} \frac{\partial Q}{\partial\beta_{yl}} &= \sum_{i=1}^{n}\Big\{y_iE\left[x_{il}|y,x^{obs},\theta^{(0)}\right] \\ &\quad -E\left[\frac{x_{il}}{1+e^{-(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})}}\Big|y_i,x_i^{obs},\theta^{(0)}\right]\Big\}, \end{aligned} \tag{6.3}$$

where $l=1,\ldots,p$. When $l\neq 1$, the first term on the right side becomes y_ix_{il}.

2. For β_{jk},

$$\begin{aligned} \frac{\partial Q}{\partial\beta_{jk}} &= -\frac{1}{\sigma_j^2}\sum_{i=1}^{n}\Big\{\beta_{jk}E\left[x_{ik}^2|y_i,x_i^{obs},\theta^{(0)}\right] \\ &\quad -E[x_{ik}(x_{ij}-\beta_{j0}-\beta_{j2}x_{i2}-\ldots-\beta_{j(k-1)}x_{i(k-1)}-\beta_{j(k+1)}x_{i(k+1)} \end{aligned}$$

$$- \ldots - \beta_{jp}x_{ip})|y_i, x_i^{obs}, \theta^{(0)}]\Big\}, \tag{6.4}$$

where $j = 1, \ldots, p-1; k = j, \ldots, p$.

The second derivatives are as follows.

1. For β_{yl},

$$\frac{\partial^2 Q}{\partial \beta_{yl}^2} = -\sum_{i=1}^{n} \left\{ E\left[x_{il}^2 \frac{e^{\mathbf{x}_i\beta}}{(1+e^{\mathbf{x}_i\beta})^2} \middle| y_i, x_i^{obs}, \theta^{(0)} \right] \right\} \tag{6.5}$$

and

$$\frac{\partial^2 Q}{\partial \beta_{yl} \partial \beta_{yl'}} = -\sum_{i=1}^{n} \left\{ E\left[x_{il}x_{il'} \frac{e^{\mathbf{x}_i\beta}}{(1+e^{\mathbf{x}_i\beta})^2} \middle| y_i, x_i^{obs}, \theta^{(0)} \right] \right\}, \tag{6.6}$$

where $l = 1, \ldots, p$, $l \neq l'$.

2. For β_{jk},

$$\frac{\partial^2 Q}{\partial \beta_{jk}^2} = -\frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ E\left[x_{ik}^2 \middle| y_i, x_i^{obs}, \theta^{(0)} \right] \right\} \tag{6.7}$$

$$\frac{\partial^2 Q}{\partial \beta_{jk} \partial \beta_{jl}} = -\frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ E\left[x_{ik}x_{il} \middle| y_i, x_i^{obs}, \theta^{(0)} \right] \right\}, \tag{6.8}$$

where $j = 1, \ldots, p-1; l, k = j, \ldots, p; l \neq k$ and

$$\frac{\partial^2 Q}{\partial \beta_{jk} \partial \beta_{j'l}} = 0 \tag{6.9}$$

when $j \neq j'$. They are the same as those from Equation 5.49 and Equation 5.51 in Chapter 4.

To compute above expectations, we propose to use GQ approximation methods because intractable integrations are involved. In the simple case with one covariate missing, we first write out the conditional density function $p(x_1|x_1,\ldots,y)$ as

$$p(x_1|x_2,\ldots,y) = \frac{p(y|x_1,\ldots,x_p)p(x_1|x_2,\ldots,x_p)}{\int p(y|x_1,\ldots,x_p)p(x_1|x_2,\ldots,x_p)dx_1}.$$

And the denominator can be calculated as follows.

$$\begin{aligned}
a_0 &= \int p(y|x_1,\ldots,x_p)p(x_1|x_2,\ldots,x_p)dx_1 \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int \frac{1}{1+\exp(-\mathbf{x}\beta)} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} dx_1 \\
&= \frac{1}{\sqrt{2\pi}}\int \frac{1}{1+\exp(-\beta_{y0}-\beta_{y1}(\mu+w\sigma)-\ldots-\beta_{yp}x_p)} e^{-\frac{w^2}{2}} dw \\
&= \frac{1}{\sqrt{\pi}}\int \frac{1}{1+\exp(-\beta_{y0}-\beta_{y1}(\mu+\sqrt{2}\sigma z)-\ldots-\beta_{yp}x_p)} e^{-z^2} dz \\
&\approx \sum_{i=0}^{N-1} \frac{1}{\sqrt{\pi}} \frac{e^{-x_i^2}}{1+\exp(-\beta_{y0}-\beta_{y1}(\mu+\sqrt{2}\sigma x_i)-\ldots-\beta_{yp}x_p)}.
\end{aligned}$$

For $x^{obs} = (x_2,\ldots,x_p)$, the conditional expectations can be computed by the Gaussian-Hermite quadratures as

$$\begin{aligned}
E[f(x_1)|x^{obs},\theta^{(0)}] &= \int f(x_1)p(x_1|x_2,\ldots,y)dx_1 \\
&= \frac{1}{a_0}\int f(x_1)p(y|x_1,\ldots,x_p)p(x_1|x_2,\ldots,x_p)dx_1 \\
&= \frac{1}{a_0\sqrt{2\pi}\sigma}\int \frac{f(x_1)}{1+exp(-\mathbf{x}\beta)} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} dx_1 \\
&= \frac{1}{a_0\sqrt{2\pi}}\int \frac{f(\mu+w\sigma)e^{-\frac{w^2}{2}}dw}{1+exp(-\beta_{y0}-\beta_{y1}(\mu+w\sigma)-\ldots-\beta_{yp}x_p)} \\
&= \frac{1}{a_0\sqrt{\pi}}\int \frac{f(\mu+\sqrt{2}\sigma z)e^{-z^2}dz}{1+exp(-\beta_{y0}-\beta_{y1}(\mu+\sqrt{2}\sigma z)-\ldots-\beta_{yp}x_p)} \\
&\approx \sum_{i=0}^{N-1} \frac{w_i}{a_0\sqrt{\pi}} \frac{f(\mu+\sqrt{2}\sigma x_i)e^{-x_i^2}}{1+exp(-\beta_{y0}-\beta_{y1}(\mu+\sqrt{2}\sigma x_i)-\ldots-\beta_{yp}x_p)}, \quad (6.10)
\end{aligned}$$

where $x_i, w_i, i = 1, \ldots, N-1$ are the abscissas and weights of the N-point Gaussian-Hermite quadrature, respectively. Once we have $\frac{\partial \mathbf{Q}}{\partial \beta}$ and $\frac{\partial^{\mathbf{2}} \mathbf{Q}}{\partial \beta^{\mathbf{2}}}$, we use Newton-Raphson to update η by

$$\eta^{(k+1)} = \eta^{(k)} - \frac{\partial Q(\theta^{(k)})/\partial \beta}{\partial^2 Q(\theta^{(k)})/\partial \beta^2}. \tag{6.11}$$

The intercept term $\hat{\beta}_{y0}$ can be updated iteratively using Newton-Raphson algorithm, in which its first and second derivatives can be obtained from Equation 6.3, Equation 6.5 and Equation 6.6 by plugging $x_{il} = 1$ when $l = 0$. Denote $\bar{\mathbf{x}}_{\mathbf{j'}} = (\frac{\sum_{i=1}^n E(x_{i(j'+1)}|y_i, x_i^{obs}, \theta^{(0)})}{n}, \ldots, \frac{\sum_{i=1}^n E(x_{ip}|y_i, x_i^{obs}, \theta^{(0)})}{n})$ and $\hat{\beta}_{j'} = (\hat{\beta}_{j'(j'+1)}, \ldots, \hat{\beta}_{j'p})$. We can update other components of $\gamma = (\sigma^2_{j'}, \beta_{y0}, \beta_{j'0}), j' = 1, \ldots, p-1$ from $\eta^{(k+1)}$ as follows.

$$\hat{\beta}_{j'0} = \frac{1}{n} \sum_{i=1}^n E\{x_{ij'}|y_i, x_i^{obs}, \theta^{(0)}\} - \hat{\beta}_{j'} \bar{\mathbf{x}}_{\mathbf{j'}} \tag{6.12}$$

and

$$\hat{\sigma}^2_{j'} = \frac{\sum_{i=1}^n \{E(x_{ij}|y_i, x_i^{obs}, \theta^{(0)}) - \hat{\beta}_{j'0} - \hat{\beta}_{j'0} E(\mathbf{x_i}|y_i, x_i^{obs}, \theta^{(0)})\}^2}{n}. \tag{6.13}$$

Iteratively solving Equation 6.11 to until it converges, we can obtain the unpenalized maximum likelihood estimator $(\hat{\eta}, \hat{\gamma})$, where η is the interest of parameter and subject to penalty. Following similar arguments in Equation 5.35 to Equation 5.37, we know that the penalized pseudo likelihood utilizing Q-function is to maximize

$$Q(\theta|\hat{\theta}) - n \sum_{j=1}^{p-1} \sum_{k=1}^{p} p_\lambda(|\beta_{jk}|). \tag{6.14}$$

From Equation 5.39, we know that when we take the fixed value of γ to be the unpenalized maximum likelihood estimate $\tilde{\gamma}$, maximizing Equation 6.14 is equivalent to minimize

$$(\eta - \tilde{\eta})^T \frac{-\ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta})}{n}(\eta - \tilde{\eta}) + \sum_{j=1}^{p-1}\sum_{k=1}^{p} p_\lambda(|\beta_{jk}|). \tag{6.15}$$

When $-\ddot{Q}_{\eta\eta}(\tilde{\theta}|\tilde{\theta})$ is positive-definite, solving Equation 6.15 is the same as to solve the penalized least square with SCAD penalty function, which can be solved by the LLA algorithm (Zou and Li, 2008) using LARS. To perform the one step algorithm to maximum penalized likelihood in section 4.4, we only need to minimize

$$\frac{1}{2}(\eta - \hat{\eta})^T \left\{ \sum_{i=1}^{n} \dot{Q}_i^T(\hat{\theta} \mid \hat{\theta}) \dot{Q}_i(\hat{\theta} \mid \hat{\theta}) \right\} (\eta - \hat{\eta}) + \sum_{j=1}^{p-1}\sum_{k=1}^{p} p_\lambda(|\beta_{jk}|) \tag{6.16}$$

with respect to η, where $\hat{\theta} = (\hat{\eta}, \hat{\gamma})$ is the unpenalized maximum likelihood estimator.

6.1.2 Logistic Regression with Two Missing Covariates

The only difference in the case when there are two covariates missing from the case with one covariate missing is that 2-dimension GQ summations are used to compute the conditional expectations. Specifically, when an observation has both covariates missing, we can calculate a_0 as follows.

$$\begin{aligned}
a_0 &= \iint p(y|x_1,\ldots,x_p)p(x_1|x_2,\ldots,x_p)dx_1\,dx_2 \\
&= \frac{1}{2\pi}\iint p(y|x_1,\ldots,x_p)\frac{\frac{1}{|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}}{e^{-x_1^2/2}e^{-x_2^2/2}}e^{-x_1^2/2}e^{-x_2^2/2}dx_1\,dx_2 \\
&= \frac{1}{2\pi|\Sigma|^{1/2}}\int\frac{\int\frac{p(y|x_1,\ldots,x_p)e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}}{e^{-x_1^2/2}}e^{-x_1^2/2}dx_1}{e^{-x_2^2/2}}e^{-x_2^2/2}dx_2.
\end{aligned}$$

As a result, above expression can be approximately by the 2-dimension Gauss-Hermite quadrature. And we can compute the conditional expectations for the first and second derivatives in a similar way.

6.1.3 A Simulation Study

In the following example, we compare the proposed variable selection method for logistic regression based on likelihood of incomplete data to the variable selection method based on imputed data. Two imputation methods are compared. One is the normal model of Schafer (1998). The other is the fully conditional specification of Burren et al. (2006). Both of the imputation methods are implemented to generate one imputed data set and maximizing penalized full data likelihood method for variable selection is applied afterwards. In generating the data, we used the same specification for the continuous covariates in the previous section of multivariate normal and add a binary dependent variable. Details are as follows.

Model 2 (Logistic Regression) We chose $\mathbf{x} \sim N(0, \Sigma)$ a $n \times p$ matrix and y a binary response. We set $n = 200, 400$ and 800 and $p = 10$. $\Sigma_{10\times 10}$ is constructed the same way in model 1a. Since we have 11 variables here, the number of coefficients in the consecutive regression models is $\frac{p(p-1)}{2} = 55$. For the coefficients in the logistic regression $\log \frac{p}{1-p} = \mathbf{x}\beta$, we only set two of them non-zeros: $\beta_{y1} = 1$ and $\beta_{y3} = 3$. Thus, the total number of non-zero coefficients is 12. They are $\beta_{y1}, \beta_{y3}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{20}, \beta_{21}, \beta_{22}, \beta_{28}, \beta_{29}$ and β_{35}. Missing Complete At Random mechanism is implemented as follows: for each variable, randomly select 20% of subjects to have missing values. We repeated the simulation for 500 times.

The simulation results are given in *Table VIII*. From *Table VIII*, it can be seen that our proposed method for model selection outperforms the two imputation methods with respect to reducing the model error and identifying the correct model. All three methods on average can identify at least 11 of 12 non-zero coefficients out of a total 55 regression coefficients.

TABLE VIII

SIMULATION RESULT FOR LOGISTIC REGRESSION WITH TWO MISSING COVARIATES

			No. of Zeros/Non-zeros		Proportion of		
Method	Sample Size	MRME	C (sd)	IC (sd)	Under-fit	Correct-fit	Over-fit
	n=200	0.534	11.7 (0.54)	1.85 (1.31)	0.29	0.17	0.54
L-func	n=400	0.438	11.8 (0.37)	0.88 (1.13)	0.17	0.54	0.29
	n=800	0.354	11.9 (0.29)	0.40 (0.83)	0.09	0.76	0.15
	n=200	0.861	11.7 (0.46)	2.39 (1.44)	0.23	0.12	0.66
norm	n=400	0.746	11.8 (0.37)	1.26 (1.19)	0.17	0.37	0.46
	n=800	0.616	11.9 (0.24)	0.83 (1.06)	0.09	0.55	0.36
	n=200	0.715	11.7 (0.54)	2.21 (1.46)	0.21	0.16	0.63
mice	n=400	0.678	11.8 (0.35)	1.33 (1.21)	0.20	0.36	0.44
	n=800	0.593	11.9 (0.27)	0.68 (1.03)	0.09	0.63	0.28

Their performances are getting better with smaller MRME and larger correct-fit proportions when sample size n increases. When sample size is small ($n = 200$) with relative large missing proportions in covariates (20%), both likelihood approach and imputation method tend to over-fit the true model. As sample size increases, likelihood approach outperforms the other methods with less overfits and more correct fits.

6.2 Logistic Regression with Arbitrary Missing Continuous Covariates

6.2.1 Monte Carlo Simulation

In practice, we often have more than one covariate subject to missing. In the following three sections, we will extend the current methods to model selection in logistic regression with arbitrary number of missingness in continuous covariates, binary covariates and in mixed covariates including both continuous and binary covariates, respectively.

When continuous covariates are subject to missing, one challenge is to evaluate intractable integrations in the first and second derivatives from Equation 6.3 to Equation 6.8.

Gauss Quadrature approximation will become less efficient in this case since it involves multiple-level summations. For example, in a logistic regression with p covariates, for a given observation i, suppose covariates x_1, x_3 and x_5 are subject to missing, to evaluate the conditional expectation of a general function g for the missing covariates given observed covariates and estimated parameters, three layers of summation are needed in GQ approximation. One alternative is to use Monte Carlo simulation method to approximate the conditional expectations in evaluating the Q-function in the EM algorithm. Specifically in our case, since the density function $f(x^{mis}|y, x^{obs}, \theta^{(0)})$ involves a logistic density and a multivariate normal density, we can use rejection sampling method along with Monte Carlo simulations in performing EM algorithm.

Let $x_i = (x_{i,m}^T, x_{i,o}^T)$. By taking a Monte Carlo sample of size K from rejection sampler from the density $f(x^{mis}|y, x^{obs}, \theta^{(0)})$: $\mathbf{x}_{i,m}^{(k)}, k = 1, \ldots, K$. Then the expectation $E[g(x_{i,m})|y, x^{obs}, \theta^{(0)}]$ can be approximated as

$$E[g(x_{i,m})|y, x^{obs}, \theta^{(0)}] \approx \frac{1}{K} \sum_{k=1}^{K} g(\mathbf{x}_{i,m}^{(s,k)}),$$

where $\mathbf{x}_{i,m}^{(s,k)}$ is the kth simulated value at the sth iteration in EM algorithm.

Monte Carlo sampling with rejection method can be implemented as follows. For each $i = 1, \ldots, n$, we can write the conditional density $f(x_{i,m}|y, x_{i,o}, \theta^{(0)})$ as

$$f(x_{i,m}|y, x_{i,o}, \theta^{(0)}) \propto f_1(y|x_{i,m}, x_{i,o}, \theta^{(0)}) * f_2(x_{i,m}|x_{i,o}, \theta^{(0)}),$$

where the first term on the right side is the logistic density and the second is a multivariate normal density. First, we sample $x_{i,m}^{(1)}$ from a multivariate normal distribution, whose mean and covariance structure can be obtained through sweep operator with $x_{i,o}$ and initial parameters

$\theta^{(0)}$. Using the sampled values of $x_{i,m}^{(1)}$ with $x_{i,o}$, we can compute f_1. Then, generate a uniformly distributed random variable on $[0,1]$ and compare it with f_1. If the random number is greater than f_1, reject $x_{i,m}^{(1)}$ and repeat the step in generating $x_{i,m}$. Otherwise accept it and thus generated a random sample following the distribution of $f(x_{i,m}|y, x_{i,o}, \theta^{(0)})$. Repeat the above steps until we get K samplers $x_{i,m}^{(k)}, k = 1, \ldots, K$. Once we get the approximations to the conditional expectations in the Q-function $Q(\theta|\theta^{(0)})$ defined in Equation 6.2, we can perform EM algorithm to get the maximum likelihood estimate $\tilde{\theta}$ and following section 4.4, the one-step maximum penalized likelihood estimate can be obtained by minimizing

$$\frac{1}{2}(\theta - \tilde{\theta})^T \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{Q}(\mathbf{x}_i^{obs}, \tilde{\theta})^T \dot{Q}(\mathbf{x}_i^{obs}, \tilde{\theta}) \right\} (\theta - \tilde{\theta}) + \sum_{j=1}^{p'} p_{\lambda_n}(|\theta_j|). \tag{6.17}$$

Simulation can be conducted similarly as the structure set up for *model 2* in which missingness can be extended to more than two continuous covariates.

6.2.2 A Simulation Study

In the following example, we compare our proposed variable selection method for logistic regression with missing covariates with the variable selection method proposed in Ibrahim et al. (2008), and the variable selection based on imputed fully data with imputation by the fully conditional specification of Burren et al. (2006) . Simulation details are as follows.

Model 2a (Continuous Covariates) We generate $\mathbf{x} \sim N(0, \Sigma)$ and y a binary response, following a logistic regression model. The sample size is set to be n=300 and 600. There are p=6 covariates. The covariance matrix $\Sigma_{6\times 6}$ is constructed as follows. Let A be a correlation matrix with pairwise correlation $0.5^{|j_1 - j_2|}$ between x_{j1} and x_{j2} for $j_1, j_2 \leq 3$; otherwise set to 0. Then let $\Sigma = A^T A$. Under this variance-covariance matrix, the non-zero covariate regression model coefficients are $\beta_{12} = 1.14, \beta_{13} = -0.71$ and $\beta_{23} = 0.80$. For the coefficients in the logistic

TABLE IX

SIMULATION RESULT FOR LOGISTIC REGRESSION WITH CONTINUOUS DATA (MAR)

			No. of Zeros/Non-zeros		Proportion of		
Method	Sample Size	MRME	C (sd)	IC (sd)	Under-fit	Correct-fit	Over-fit
	n=300	0.73	4.93 (0.26)	0.43 (0.73)	0.08	0.70	0.22
L-func	n=600	0.63	5.00 (0.07)	0.12 (0.36)	0.01	0.90	0.09
	n=300	0.79	4.94 (0.24)	0.48 (0.81)	0.06	0.70	0.24
Q-func	n=600	0.67	4.99 (0.10)	0.18 (0.50)	0.02	0.85	0.13
	n=300	0.78	4.92 (0.27)	0.81 (1.02)	0.08	0.54	0.38
mice	n=600	0.68	4.97 (0.16)	0.58 (0.89)	0.03	0.64	0.33

regression $\log \frac{P(y=1)}{1-P(y=1)} = \mathbf{x}\beta$, we only set two of them to be non-zeros: $\beta_{y1} = 1$ and $\beta_{y3} = 3$. Since we have 7 variables, the number of consecutive regression coefficients is $\frac{p(p-1)}{2} = 21$. And the total number of non-zero coefficients is 5. They are $\beta_{y1}, \beta_{y3}, \beta_{12}, \beta_{13}$ and β_{23}. Missing At Random mechanism is implemented as follows. Covariates x_1 and x_6 are complete. The probabilities that covariates $x_w, w = 2, 3, 4, 5$ are subject to missing are set to be

$$expit(\alpha_w + \gamma_1 x_{i1} + \gamma_2 x_{i6}),$$

where $expit(u) = \frac{\exp(u)}{1+\exp(u)}$, $\alpha_2 = 1.7, \alpha_3 = 1.6, \alpha_4 = 1.5, \alpha_5 = 1.5$, $\gamma_1 = 0.8$ and $\gamma_2 = -0.3$. The average missing proportions for $x_w, w = 2, 3, 4, 5$ are $12.0\%, 15.1\%, 13.0\%$ and 16.7%, respectively. The simulation were repeated for 200 times.

Table IX shows the simulation results. It can be seen that the proposed model selection method identifies 90% of the correct model, compared with 85% and 64% identified by the penalized EM algorithm proposed in Ibrahim et al. (2008), and the imputation algorithm in Burren et al. (2006), respectively. The proposed penalized likelihood approach has the

TABLE X

MEANS OF REGRESSION COEFFICIENTS IN LOGISTIC REGRESSION WITH CONTINUOUS DATA

Method	Coefficient	n=300	n=600	true value
	β_1	0.86 (0.26)	0.89 (0.13)	1.00
	β_2	2.87 (0.18)	2.86 (0.14)	3.00
L-func	β_3	1.13 (0.06)	1.14 (0.03)	1.14
	β_4	-0.66 (0.05)	-0.71 (0.03)	-0.71
	β_5	0.80 (0.04)	0.80 (0.03)	0.80
	β_1	0.85 (0.24)	0.88 (0.13)	1.00
	β_2	2.91 (0.19)	2.87 (0.13)	3.00
Q-func	β_3	1.14 (0.04)	1.14 (0.03)	1.14
	β_4	-0.71 (0.04)	-0.71 (0.03)	-0.71
	β_5	0.80 (0.04)	0.80 (0.03)	0.80
	β_1	0.94 (0.46)	0.98 (0.28)	1.00
	β_2	2.86 (0.42)	2.91 (0.28)	3.00
mice	β_3	1.11 (0.09)	1.09 (0.07)	1.14
	β_4	-0.68 (0.18)	-0.62 (0.31)	-0.71
	β_5	0.78 (0.05)	0.80 (0.04)	0.80

smallest model error among the other competitors and has the highest probability to identity true model. When sample size $n = 300$, all three algorithms misclassify rates of 0.43, 0.48, and 0.81 zero coefficient to non-zero coefficient respectively, causing the over-fit. As sample size increases, the proposed algorithm reduces this error by the most to 0.12. For the column "C," the proposed method based on observed likelihood improves its performance the most in selecting all 5 non-zero coefficients to be non-zeros when sample size increases from 300 to 600, compared with the other two selection algorithms. As we see, the proposed penalized likelihood approach improves significant as the sample size increases. *Table X* lists the means of all the non-zero regression coefficients. It can be seen that all three methods give pretty close estimates to their corresponding truth values.

6.3 Logistic Regression with Arbitrary Missing Binary Covariates

6.3.1 Computation in Expectation-Maximization Algorithm

In this section, we extend our methodology of variable selection to the logistic regression with arbitrary missingness to the case where all covariates are binary. Suppose the outcome variable y and p covariates are all binary. The observations are $(y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n$. We model the outcome variable and covariates as follows.

$y|x_1, x_2, \ldots, x_p \sim Bernoulli\{p(\beta_{y0} + \beta_{y1}x_1 + \ldots + \beta_{yp}x_p)\}$

$x_1|x_2, \ldots, x_p \sim Bernoulli\{p(\beta_{10} + \beta_{1,2}x_2 + \ldots + \beta_{1,p}x_p)\}$

$x_2|x_3, \ldots, x_p \sim Bernoulli\{p(\beta_{20} + \beta_{2,3}x_3 + \ldots + \beta_{2,p}x_p)\}$

$\ldots$

$x_{p-1}|x_p \sim Bernoulli\{p(\beta_{(p-1)0} + \beta_{p-1,p}x_p)\},$

where $p(u) = \frac{e^u}{1+e^u}$.

Thus, the full data log-likelihood is

$$\begin{aligned} l(\beta) \quad = \quad & \sum_{i=1}^{n}\{y_i(\beta_{y,0} + \beta_{y,1}x_{i1} + \ldots + \beta_{y,p}x_{ip}) \\ & - \log[1 + e^{\beta_{y,0}+\beta_{y,1}x_{i1}+\ldots+\beta_{y,p}x_{ip}}] \\ & + x_{i1}(\beta_{1,0} + \beta_{1,2}x_{i2} + \ldots + \beta_{1,p}x_{ip}) \\ & - \log[1 + e^{\beta_{1,0}+\beta_{1,2}x_{i2}+\ldots+\beta_{1,p}x_{ip}}] \\ & + x_{i2}(\beta_{2,0} + \beta_{2,3}x_{i3} + \ldots + \beta_{2,p}x_{ip}) \\ & - \log[1 + e^{\beta_{2,0}+\beta_{2,3}x_{i3}+\ldots+\beta_{2,p}x_{ip}}] \\ & \ldots\ldots \\ & + x_{i(p-1)}(\beta_{p-1,0} + \beta_{p-1,p}x_{ip}) \\ & - \log[1 + e^{\beta_{p-1,0}+\beta_{p-1,p}x_{ip}}]\}. \end{aligned} \tag{6.18}$$

We then write the Q-function in EM algorithm as,

$$\begin{aligned}
Q(\theta|(y, x^{obs}, \theta^{(0)})) &= \sum_{i=1}^{n}\{y_i(\beta_{y,0} + \beta_{y,1}E[x_{i1}|y, x^{obs}, \theta^{(0)}] + \ldots + \beta_{y,p}E[x_{ip}|y, x^{obs}, \theta^{(0)}]) \\
&\quad - E\{\log[1 + e^{\beta_{y,0}+\beta_{y,1}x_{i1}+\ldots+\beta_{y,p}x_{ip}}]|y, x^{obs}, \theta^{(0)}\} \\
&\quad + (\beta_{1,0}E[x_{i1}|y, x^{obs}, \theta^{(0)}] + \beta_{1,2}E[x_{i1}x_{i2}|y, x^{obs}, \theta^{(0)}] \\
&\quad + \ldots + \beta_{1,p}E[x_{i1}x_{ip}|y, x^{(obs)}, \theta^{(0)}]) \\
&\quad - E\{\log[1 + e^{\beta_{1,0}+\beta_{1,2}x_{i2}+\ldots+\beta_{1,p}x_{ip}}]|y, x^{obs}, \theta^{(0)}\} \\
&\quad + (\beta_{2,0}E[x_{i2}|y, x^{obs}, \theta^{(0)}] + \beta_{2,3}E[x_{i2}x_{i3}|y, x^{obs}, \theta^{(0)}] \\
&\quad + \ldots + \beta_{2,p}E[x_{i2}x_{ip}|y, x^{(obs)}, \theta^{(0)}]) \\
&\quad - E\{\log[1 + e^{\beta_{2,0}+\beta_{2,3}x_{i3}+\ldots+\beta_{2,p}x_{ip}}]|y, x^{obs}, \theta^{(0)}\} \\
&\quad \ldots\ldots \\
&\quad + (\beta_{p-1,0}E[x_{i(p-1)}|y, x^{obs}, \theta^{(0)}] + \beta_{p-1,p}E[x_{i(p-1)}x_{ip}|y, x^{obs}, \theta^{(0)}]) \\
&\quad - E\{\log[1 + e^{\beta_{p-1,0}+\beta_{p-1,p}x_{ip}}]|y, x^{obs}, \theta^{(0)}\}\}, \qquad (6.19)
\end{aligned}$$

where $\theta = (\beta_{y,k}, \beta_{j,k}, \beta_{y,0}, \beta_{j,0})$, $j = 1, \ldots, p-1; k = j, \ldots, p$.

The first and second derivatives with respect to the non-intercept coefficients in θ are as follow.

1. For β_{yk},

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_{yk}} &= \sum_{i=1}^{n}\left\{y_i E\left[x_{ik}\middle|y, x^{obs}, \theta^{(0)}\right]\right. \\
&\quad \left. - E\left[\frac{x_{ik}}{1 + e^{-(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})}}\middle|y, x^{obs}, \theta^{(0)}\right]\right\}, \qquad (6.20)
\end{aligned}$$

where k=1, $\ldots$, p.

2. For β_{jk},

$$\frac{\partial Q}{\partial \beta_{jk}} = \sum_{i=1}^{n} \Big\{ E\left[x_{ij}x_{ik} \Big| y, x^{obs}, \theta^{(0)} \right] - E\left[\frac{x_{ik}}{1+e^{-(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})}} \Big| y, x^{obs}, \theta^{(0)} \right] \Big\}, \quad (6.21)$$

where $j = 1, \ldots, p-1; k = j+1, \ldots, p$.

Second derivatives are written as follows.

1. For β_{yk},

$$\frac{\partial^2 Q}{\partial \beta_{yk}^2} = -\sum_{i=1}^{n} E\left[x_{ik}^2 \frac{e^{(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})}}{(1+e^{(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})})^2} \Big| y_i, x_i^{obs}, \theta^{(0)} \right]. \quad (6.22)$$

$$\frac{\partial^2 Q}{\partial \beta_{yk}\partial \beta_{yl}} = -\sum_{i=1}^{n} E\left[x_{ik}x_{il} \frac{e^{(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})}}{(1+e^{(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})})^2} \Big| y_i, x_i^{obs}, \theta^{(0)} \right], \quad (6.23)$$

where $k = 1, \ldots, p$.

2. For β_{jk},

$$\frac{\partial^2 Q}{\partial \beta_{jk}^2} = -\sum_{i=1}^{n} E\left[x_{ik}^2 \frac{e^{(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})}}{(1+e^{(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})})^2} \Big| y_i, x_i^{obs}, \theta^{(0)} \right]. \quad (6.24)$$

$$\frac{\partial^2 Q}{\partial \beta_{jk}\partial \beta_{jl}} = -\sum_{i=1}^{n} E\left[x_{ik}x_{il} \frac{e^{(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})}}{(1+e^{(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})})^2} \Big| y_i, x_i^{obs}, \theta^{(0)} \right], \quad (6.25)$$

$$\frac{\partial^2 Q}{\partial \beta_{jk}\partial \beta_{j'l}} = 0, \quad (6.26)$$

where $j = 1, \ldots, p-1; k = j+1, \ldots, p$, $l = j+1, \ldots, p; l \neq k$ and $j \neq j'$.

Since all covariates are binary, we can use summation over missing patterns to compute the expectation terms in the above calculated derivatives. Specifically,

$$\begin{aligned} & E[x_{ik}|y, x^{obs}, \theta^{(0)}] = P(x_{ik} = 1|y, x^{obs}, \theta^{(0)}] \\ = & \frac{\sum_{x^{mis}} P(y|x_1, \ldots, x_p, x_k = 1, \theta^{(0)}) P(x_1|x_2, \ldots, x_p, x_k = 1, \theta^{(0)}) \ldots P(x_{p-1}|x_p, x_k = 1, \theta^{(0)})}{\sum_{x^{mis}} P(y|x_1, \ldots, x_p, \theta^{(0)}) P(x_1|x_2, \ldots, x_p, \theta^{(0)}) \ldots P(x_{p-1}|x_p, \theta^{(0)})}, \end{aligned}$$

where each term in numerator and denominator is a logistic regression and the summation $\sum_{x^{mis}}$ is over all possible values of x^{mis}. To compute the expectation of a general function $g(\cdot)$ of x^{mis} given y, x^{obs} and previous model parameter $\theta^{(0)}$, we have

$$E[g(x^{mis})|y, x^{obs}, \theta^{(0)}] = \sum_{x^{mis}}^{S} g(x^{mis} = s) P(x^{mis} = s|y, x^{obs}, \theta^{(0)}),$$

where s is one of the patterns of x^{mis} and S is the number of all patterns. For an observation $x_{i1}, x_{i2}, \ldots, x_{ip}$, if q variables are subject to missing, then $S = 2^q$. Similarly,

$$P(x^{mis} = s|y, x^{obs}, \theta^{(0)}) = \frac{P(y, x^{obs}, x^{mis} = s|\theta^{(0)})}{P(y, x^{obs}|\theta^{(0)})},$$

where the denominator can be computed through summation over all possible values of x^{mis} on the join distribution of $(y, x|\theta^{(0)})$, which is written as the product of consecutive condition density functions.

6.3.2 A Simulation Study

In this section, comparison is made between the proposed variable selection algorithm in logistic regression with missing binary covariates data and the method proposed in Ibrahim

et al. (2008), with the selection criterion described in Ibrahim et al. (2008). Simulation details are as follow.

Model 2b (Binary Covariates) Let $\mathbf{x}$ be a $n \times p$ matrix and y a binary response. The sample size is set to be n=300 and 600. There are p=6 covariates. We set the consecutive regression models for covariates and outcome variable as follows. Let $x_6 \sim Bernoulli(0.6)$; $x_5|x_6 \sim Bernoulli(p(-0.8x_6))$; $x_4|x_5, x_6 \sim Bernoulli(p(2x_5))$; $x_3|x_4, x_5, x_6 \sim Bernoulli(0.5))$; $x_2|x_3, x_4, x_5, x_6 \sim Bernoulli(p(1.2x_4))$; $x_1|x_2, x_3, x_4, x_5, x_6 \sim Bernoulli(p(x_6))$ and $y|x_1, x_2, x_3, x_4, x_5, x_6 \sim Bernoulli(p(2x_4))$. Total number of consecutive regression coefficients is $\frac{p(p+1)}{2} = 21$. And the number of non-zero coefficients is 5. They are $\beta_{y4} = 2.0, \beta_{16} = 1.0, \beta_{24} = 1.2, \beta_{45} = 2.0$ and $\beta_{56} = -0.8$. Missing At Random(MAR) mechanism is implemented as follows. Only covariate x_6 is set to be complete. The probability that covariates $x_w, w = 1, 2, 3, 4, 5$ are subject to missing is set to be

$$expit(\alpha_w + \gamma_1 y_i + \gamma_2 x_{i6}),$$

where $expit(u) = \frac{exp(u)}{1+exp(u)}$, $\alpha_1 = 1.2, \alpha_2 = 1.3, \alpha_3 = 1.1, \alpha_4 = 1.0, \alpha_5 = 0.9$, $\gamma_1 = 0.8$ and $\gamma_2 = -0.3$. The average missing proportions for $x_w, w = 1, 2, 3, 4, 5$ are $17.0\%, 13.3\%, 16.8\%, 18.8\%$ and 21.5%, respectively. We run the simulation for replicates of 200.

For the case with all binary covariates, it can be seen from *Table XI* that the proposed method identifies 71% of the correct model, compared with 60% and 55% for the penalized EM algorithm proposed in Garcia et al. (2010), and the imputation algorithm based on fully conditional specifications in Burrent et al. (2006), respectively. On average, three algorithms correctly select 4.54, 4.61, and 4.66 non-zero coefficients to be non-zeros, respectively and mis-classify 1.10, 1.74, and 1.04 zero coefficients as non-zero, respectively. The column "C"

TABLE XI

SIMULATION RESULT FOR LOGISTIC REGRESSION WITH BINARY COVARIATES (MAR)

			No. of Zeros/Non-zeros		Proportion of		
Method	Sample Size	MRME	C (sd)	IC (sd)	Under-fit	Correct-fit	Over-fit
	n=300	0.59	4.54 (0.63)	1.10 (1.11)	0.29	0.40	0.31
L-func	n=600	0.44	4.86 (0.36)	0.29 (0.57)	0.08	0.71	0.21
	n=300	0.76	4.61 (0.54)	1.74 (1.38)	0.33	0.34	0.33
Q-func	n=600	0.56	4.78 (0.50)	0.59 (0.79)	0.14	0.60	0.26
	n=300	0.76	4.66 (0.55)	1.04 (0.92)	0.30	0.33	0.37
mice	n=600	0.47	4.69 (0.58)	0.27 (0.49)	0.20	0.55	0.25

TABLE XII

MEANS OF REGRESSION COEFFICIENTS IN LOGISTIC REGRESSION WITH BINARY COVARIATES

Method	Coefficient	n=300	n=600	true value
	β_1	1.96 (0.46)	2.02 (0.31)	2.00
	β_2	0.87 (0.49)	0.92 (0.33)	1.00
L-func	β_3	1.18 (0.55)	1.21 (0.32)	1.20
	β_4	2.03 (0.52)	1.97 (0.33)	2.00
	β_5	-0.86 (0.39)	-0.76 (0.37)	-0.80
	β_1	2.07 (0.45)	2.00 (0.29)	2.00
	β_2	0.94 (0.45)	0.95 (0.33)	1.00
Q-func	β_3	1.12 (0.47)	1.22 (0.28)	1.20
	β_4	1.96 (0.41)	1.95 (0.34)	2.00
	β_5	-0.86 (0.37)	-0.80 (0.27)	-0.80
	β_1	2.02 (0.48)	1.98 (0.34)	2.00
	β_2	0.97 (0.45)	0.87 (0.42)	1.00
mice	β_3	1.12 (0.45)	1.17 (0.27)	1.20
	β_4	2.02 (0.54)	1.96 (0.35)	2.00
	β_5	-0.82 (0.36)	-0.77 (0.34)	-0.80

refers to under-fit effect and column "IC" refers to over-fit effect in the model selection process. Simulation result shows that our proposed penalized likelihood approach has the smallest model error among the other competitors and has the highest probability to identity true model. As can be seen, the probability to identify the correct model is much lower compared with the case with all continuous covariates due to more covariates are subject to missing with larger missing proportions. But as sample size increases, performance of model selection with penalized likelihood increases significantly (71%) in proportions of the correct fits, compared with 60% and 55% for the other two methods. Thus, the proposed model selection algorithm outperforms the other two competitive algorithms with a reasonable large sample size.

Table XII lists the means of all the non-zero regression coefficients. It can be seen that all three methods give estimates very close to their corresponding truth values.

6.4 Logistic Regression with Arbitrary Missing Mixed Continuous and Binary Covariates

6.4.1 Computation Issue in Expectation-Maximization Algorithm

When both continuous and binary covariates are subject to missing, let the outcome variable y and the first p_1 covariates be all binary, the rest $p_2 = p - p_1$ covariates be continuous and have a normal distribution $N(\mu, \Sigma)$. The full data are $(y_i, x_{i1}, \ldots, x_{ip_1}, x_{i(p_1+1)}, \ldots, x_{ip})$, $i = 1, \ldots, n$. We model the outcome variable and covariates as follows.

$y|x_1, x_2, \ldots, x_p \sim Bernoulli\{p(\beta_{y0} + \beta_{y1}x_1 + \ldots + \beta_{yp}x_p)\}$

$x_1|x_2, \ldots, x_p \sim Bernoulli\{p(\beta_{10} + \beta_{1,2}x_2 + \ldots + \beta_{1,p}x_p)\}$

$\ldots$

$x_{p_1}|x_{p_1+1}, \ldots, x_p \sim Bernoulli\{p(\beta_{(p-1)0} + \ldots + \beta_{p-1,p}x_p)\}$

$x_{p_1+1}|x_{p_1+2}, \ldots, x_p \sim N(\eta_1, \sigma_1^2)$

$x_{p_1+2}|x_{p_1+3}, \ldots, x_p \sim N(\eta_2, \sigma_2^2)$

...

$x_{p-1}|x_p \sim N(\eta_{p_2}, \sigma^2_{p_2})$,

where $p(u) = \frac{e^u}{1+e^u}$ and $\eta_k = \beta_{k0} + \beta_{k,(k+1)}x_{k+1} + \ldots + \beta_{k,p}x_p$, $k = p_1 + 1, \ldots, p-1$.

Thus, the full data log-likelihood is

$$
\begin{aligned}
l(\beta,\sigma) \;=\; & \sum_{i=1}^{n}\{y_i(\beta_{y,0} + \beta_{y,1}x_{i1} + \ldots + \beta_{y,p}x_{ip}) \\
& - \log[1 + e^{\beta_{y,0}+\beta_{y,1}x_{i1}+\ldots+\beta_{y,p}x_{ip}}] \\
& + x_{i1}(\beta_{1,0} + \beta_{1,2}x_{i2} + \ldots + \beta_{1,p}x_{ip}) \\
& - \log[1 + e^{\beta_{1,0}+\beta_{1,2}x_{i2}+\ldots+\beta_{1,p}x_{ip}}] \\
& \ldots\ldots \\
& + x_{ip_1}(\beta_{p_1,0} + \beta_{p_1,p_1+1}x_{i(p_1+1)} + \ldots + \beta_{p_1,p}x_{ip}) \\
& - \log[1 + e^{\beta_{p_1,0}+\beta_{p_1,p_1+1}x_{i(p_1+1)}+\ldots+\beta_{p_1,p}x_{ip}}]\} \\
& - \frac{1}{2}log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_{i(p_1+1)} - (\beta_{(p_1+1),0} + \beta_{(p_1+1),(p_1+2)}x_{i(p_1+2)} + \ldots + \beta_{(p_1+1),p}x_{ip}))^2 \\
& - \frac{1}{2}log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_{i(p_1+2)} - (\beta_{(p_1+2),0} + \beta_{(p_1+2),(p_1+3)}x_{i(p_1+3)} + \ldots + \beta_{(p_1+2),p}x_{ip}))^2 \\
& \ldots\ldots \\
& - \frac{1}{2}log(2\pi\sigma_{p_2-1}^2) - \frac{1}{2\sigma_{p_2-1}^2}(x_{i(p-1)} - (\beta_{(p-1)0} + \beta_{(p-1)p}x_{pi}))^2 \\
& - \frac{1}{2}log(2\pi\sigma_{p_2}^2) - \frac{1}{2\sigma_{p_2}^2}(x_{ip} - \eta_{p_2})^2\}
\end{aligned}
\tag{6.27}
$$

The corresponding Q-function in EM algorithm is as follows.

$$
\begin{aligned}
Q(\theta|(y, x^{obs}, \theta^{(0)})) \;=\; & \sum_{i=1}^{n}\{y_i(\beta_{y,0} + \beta_{y,1}E[x_{i1}|y_i, x_i^{obs}, \theta^{(0)}] + \ldots + \beta_{y,p}E[x_{ip}|y_i, x_i^{obs}, \theta^{(0)}]) \\
& - E\{\log[1 + e^{\beta_{y,0}+\beta_{y,1}x_{i1}+\ldots+\beta_{y,p}x_{ip}}]|y_i, x_i^{obs}, \theta^{(0)}\} \\
& + (\beta_{1,0}E[x_{i1}|y, x^{obs}, \theta^{(0)}] + \beta_{1,2}E[x_{i1}x_{i2}|y_i, x_i^{obs}, \theta^{(0)}]
\end{aligned}
$$

$$
\begin{aligned}
&+ \ldots + \beta_{1,p} E[x_{i1} x_{ip} | y_i, x_i^{obs}, \theta^{(0)}]) \\
&- E\{\log[1 + e^{\beta_{1,0} + \beta_{1,2} x_{i2} + \ldots + \beta_{1,p} x_{ip}}] | y_i, x_i^{obs}, \theta^{(0)}\} \\
&\ldots\ldots \\
&+ (\beta_{p_1,0} E[x_{ip_1} | y, x^{obs}, \theta^{(0)}] + \beta_{p_1,(p_1+1)} E[x_{ip_1} x_{i(p_1+1)} | y_i, x_i^{obs}, \theta^{(0)}]) \\
&+ \ldots + \beta_{p_1,p} E[x_{ip_1} x_{ip} | y_i, x_i^{obs}, \theta^{(0)}]) \\
&- E\{\log[1 + e^{\beta_{p-1,0} + \beta_{p-1,p} x_{ip}}] | y_i, x_i^{obs}, \theta^{(0)}\} \\
&- \frac{E[(x_{i(p_1+1)} - \beta_{(p_1+1),0} - \ldots - \beta_{(p_1+1),p} x_{ip})^2 | y_i, x_i^{obs}, \theta^{(0)}]}{2\sigma_1^2} \\
&\ldots\ldots \\
&- \frac{E[(x_{i(p-1)} - (\beta_{(p-1)0} + \beta_{(p-1)p} x_{pi}))^2 | y_i, x_i^{obs}, \theta^{(0)}]}{2\sigma_{p_2-1}^2} \\
&- \frac{E[(x_{ip} - \eta_{p_2})^2 | y_i, x_i^{obs}, \theta^{(0)}]}{2\sigma_{p_2}^2}\}, \qquad (6.28)
\end{aligned}
$$

where $\theta = (\beta_{y,k}, \beta_{j,k}, \beta_{y,0}, \beta_{j,0},\ j = 1, \ldots, p-1; k = j, \ldots, p,\ \sigma_1, \ldots, \sigma_{p2})$.

The first and second derivatives with respect to the non-intercept coefficients in θ are as follows.

1. For β_{yk},

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_{yk}} &= \sum_{i=1}^{n} \Big\{ y_i E\left[x_{ik} \Big| y_i, x_i^{obs}, \theta^{(0)} \right] \\
&\quad - E\left[\frac{x_{ik}}{1 + e^{-(\beta_{y0} + \beta_{y1} x_{i1} + \ldots + \beta_{yp} x_{ip})}} \Big| y_i, x_i^{obs}, \theta^{(0)} \right] \Big\}, \qquad (6.29)
\end{aligned}
$$

   where $k = 1, \ldots, p$.

2. For the coefficients β_{jk} in logistic regression models of first p_1 covariates,

$$
\frac{\partial Q}{\partial \beta_{jk}} = \sum_{i=1}^{n} \Big\{ E\left[x_{ij} x_{ik} \Big| y_i, x_i^{obs}, \theta^{(0)} \right]
$$

$$- E\left[\frac{x_{ik}}{1+e^{-(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})}}\middle| y_i, x_i^{obs}, \theta^{(0)}\right]\Bigg\}, \tag{6.30}$$

where $j = 1, \ldots, p_1; k = j+1, \ldots, p$

3. For the coefficients β_{jk} in linear regression models of $p - p_1$ covariates,

$$\begin{aligned}\frac{\partial Q}{\partial \beta_{jk}} &= \frac{1}{\sigma^2_{j-p_1}}\sum_{i=1}^{n}\Bigg\{E\left[x_{ik}x_{ij}\middle| y_i, x_i^{obs}, \theta^{(0)}\right] \\ &\quad - E\left[x_{ik}(\beta_{j0}+\beta_{j,(j+1)}x_{i(j+1)}+\ldots+\beta_{j,p}x_{ip})\middle| y_i, x_i^{obs}, \theta^{(0)}\right]\Bigg\}, \end{aligned}\tag{6.31}$$

where $j = p_1+1, \ldots, p-1; k = j+1, \ldots, p$

Second derivatives are written as follows.

1. For β_{yk},

$$\frac{\partial^2 Q}{\partial \beta^2_{yk}} = -\sum_{i=1}^{n} E\left[x_{ik}^2\frac{e^{(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})}}{(1+e^{(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})})^2}\middle| y_i, x_i^{obs}, \theta^{(0)}\right]. \tag{6.32}$$

$$\frac{\partial^2 Q}{\partial \beta_{yk}\partial \beta_{yl}} = -\sum_{i=1}^{n} E\left[x_{ik}x_{il}\frac{e^{(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})}}{(1+e^{(\beta_{y0}+\beta_{y1}x_{i1}+\ldots+\beta_{yp}x_{ip})})^2}\middle| y_i, x_i^{obs}, \theta^{(0)}\right], \tag{6.33}$$

where $k, l = 1, \ldots, p$, $k \neq l$.

2. For the coefficients β_{jk} in logistic regression models of first p_1 covariates,

$$\frac{\partial^2 Q}{\partial \beta^2_{jk}} = -\sum_{i=1}^{n} E\left[x_{ik}^2\frac{e^{(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})}}{(1+e^{(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})})^2}\middle| y_i, x_i^{obs}, \theta^{(0)}\right]. \tag{6.34}$$

$$\frac{\partial^2 Q}{\partial \beta_{jk}\partial \beta_{jl}} = -\sum_{i=1}^{n} E\left[x_{ik}x_{il}\frac{e^{(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})}}{(1+e^{(\beta_{j0}+\beta_{j(j+1)}x_{i(j+1)}+\ldots+\beta_{jp}x_{ip})})^2}\middle| y_i, x_i^{obs}, \theta^{(0)}\right], \tag{6.35}$$

$$\frac{\partial^2 Q}{\partial \beta_{jk} \partial \beta_{j'l}} = 0, \tag{6.36}$$

where $j = 1, \ldots, p_1; k = j+1, \ldots, p$, $l = j+1, \ldots, p, l \neq k$ and $j \neq j'$.

3. For the coefficients β_{jk} in linear regression models of $p - p_1$ covariates,

$$\frac{\partial^2 Q}{\partial \beta_{jk}^2} = -\frac{1}{\sigma_{j-p_1}^2} \sum_{i=1}^{n} E\left[x_{ik}^2 \middle| y, x^{obs}, \theta^{(0)}\right] \tag{6.37}$$

$$\frac{\partial^2 Q}{\partial \beta_{jk} \partial \beta_{jl}} = -\frac{1}{\sigma_{j-p_1}^2} \sum_{i=1}^{n} E\left[x_{ik} x_{il} \middle| y, x^{obs}, \theta^{(0)}\right] \tag{6.38}$$

$$\frac{\partial^2 Q}{\partial \beta_{jk} \partial \beta_{j'l}} = 0, \tag{6.39}$$

where $j = p_1 + 1, \ldots, p-1; k = j+1, \ldots, p$, $l = j+1, \ldots, p, l \neq k$ and $j \neq j'$.

In order to compute the expectation terms of the first and second derivatives in the EM algorithm, we use Monte Carlo simulations along with rejection sampling method. Details are described as follow. For each observation $i, i = 1, \ldots, n$, if only the binary covariates are subject to missing, then we can use the same technique in section 6.3. Otherwise, we will first draw a sample of missing continuous covariates then compute the expectations conditional on each of the sample values and other observed covariates. Suppose we have data $y, x_1, \ldots, x_{p1}, x_{p_1+1}, \ldots, x_p$, where y and first p_1 variables are binary and the rest are continuous.

Let x_c^{mis} and x_b^{mis} denote the missing continuous and binary covariates, respectively and x_c^{obs} and x_b^{obs} denote the observed continuous and binary covariates, respectively. Then we have

$$\begin{aligned} f(x_c^{mis}|y, x^{obs}, \theta^{(0)}) &\propto \sum_{x_b^{mis}} f_1(y, x^{obs}, x_c^{mis}, x_b^{mis}|\theta^{(0)}) \\ &= \sum_{x_b^{mis}} f_1(y, x_b^{obs}, x_b^{mis}|x_c^{mis}, x_c^{obs}, \theta^{(0)}) * f_2(x_c^{obs}, x_c^{mis}|\theta^{(0)}) \\ &= f_2(x_c^{obs}, x_c^{mis}|\theta^{(0)}) * \sum_{x_b^{mis}} f_1(y, x_b^{obs}, x_b^{mis}|x_c^{mis}, x_c^{obs}, \theta^{(0)}), \end{aligned}$$

where the second term is a summation of products of a sequential logistic regression probabilities. Thus, we can perform the rejection sampling similar to the case with all continuous covariates. We first sample x_c^{mis} from a multivariate normal distribution $f_2(x_c^{obs}, x_c^{mis}|\theta^{(0)})$. Utilizing the sampled values of x_c^{mis} to compute f_1. Then, generate a uniformly distributed random variable on $[0, 1]$ and compare it with f_1. If the random number is greater than f_1, reject x_c^{mis} and repeat the step in generating x_c^{mis}. Otherwise accept it and thus generated a random sample following the distribution of $f(x_c^{mis}|y, x^{obs}, \theta^{(0)})$. Repeat the above steps until we get the K number of rejection samplers $x_c^{mis(k)}, k = 1, \ldots, K$.

6.4.2 A Simulation Study

Similar to the previous two sections of simulation studies, in this section, the proposed variable selection method for logistic regression with missing covariates data in both continuous and binary variables is compared with the method proposed in Ibrahim et al. (2008), with the selection criterion described in Ibrahim et al. (2008), and the imputation algorithm of fully conditional specification (FCS) in Buuren et al. (2006). For method of imputation by FCS implemented under the Multivariate Imputation by Chained Equations (MICE) package in R, the dataset with missing covariates is imputed for one time then applied to model selection

method of maximizing penalized likelihood with SCAD for complete dataset. Simulation details are as follows.

Model 2c (Mixed Covariates) Let $\mathbf{x}$ be a $n \times p$ matrix and y is a binary response. There are p=6 covariates. First four variables in $\mathbf{x}$ are binary and the rest are continuous with normal distribution with mean zero and covariance Σ, where

$$\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}.$$

The sample size is set to be n=300 and 600. The binary covariates are modeled as follows. Let $x_4|x_5, x_6 \sim Bernoulli(2x_5)$; $x_3|x_4, x_5, x_6 \sim Bernoulli(0.5)$; $x_2|x_3, x_4, x_5, x_6 \sim Bernoulli(p(1.2x_4))$; $x_1|x_2, x_3, x_4, x_5, x_6 \sim Bernoulli(p(x_6))$ and $y|x_1, x_2, x_3, x_4, x_5, x_6 \sim Bernoulli(p(2x_4))$. Total number of consecutive regression model coefficients is $\frac{p(p+1)}{2} = 21$. Total number of non-zero coefficients is 5. They are $\beta_{y4} = 2, \beta_{16} = 1, \beta_{24} = 1.2, \beta_{45} = 2$ and $\beta_{56} = -0.8$. Missing At Random mechanism is implemented as follows. Covariate x_1 and x_6 are set to be completely observed. The probability that covariates $x_w, w = 2, 3, 4, 5$ are subject to missing is set to be

$$expit(\alpha_w + \gamma_1 y_i + \gamma_2 x_{i6}),$$

where $expit(u) = \frac{exp(u)}{1+exp(u)}$, $\alpha_2 = 1.7, \alpha_3 = 1.6, \alpha_4 = 1.5, \alpha_5 = 1.5$, $\gamma_1 = 0.8$ and $\gamma_2 = -0.3$. The average missing proportions for $x_w, w = 2, 3, 4, 5$ are 11.5%, 12.8%, 13.8%, and 15%, respectively. We run the simulation with of 200 replicates.

Table XIII lists simulation results. From the table, we see that the proposed method identifies 91% of the correct model, compared with 88% and 76% for penalized EM algorithm

TABLE XIII

SIMULATION RESULT FOR LOGISTIC REGRESSION WITH MIXED BINARY AND CONTINUOUS COVARIATES

			No. of Zeros/Non-zeros		Proportion of		
Method	Sample Size	MRME	C (sd)	IC (sd)	Under-fit	Correct-fit	Over-fit
	n=300	0.48	4.94 (0.24)	0.47 (0.74)	0.06	0.61	0.33
L-func	n=600	0.27	4.97 (0.17)	0.07 (0.25)	0.03	0.91	0.06
	n=300	0.45	4.86 (0.37)	0.44 (0.67)	0.14	0.55	0.31
Q-func	n=600	0.31	4.98 (0.16)	0.10 (0.32)	0.03	0.88	0.09
	n=300	0.49	4.90 (0.30)	0.62 (0.76)	0.10	0.48	0.42
mice	n=600	0.28	4.96 (0.21)	0.17 (0.31)	0.03	0.76	0.21

TABLE XIV

MEANS OF REGRESSION COEFFICIENTS IN LOGISTIC REGRESSION WITH MIXED BINARY AND CONTINUOUS COVARIATES

Method	Coefficient	n=300	n=600	true value
	β_1	2.03 (0.46)	2.03 (0.30)	2.00
	β_2	0.96 (0.27)	0.99 (0.17)	1.00
L-func	β_3	1.22 (0.38)	1.18 (0.29)	1.20
	β_4	1.95 (0.33)	1.95 (0.23)	2.00
	β_5	-0.80 (0.04)	-0.80 (0.03)	-0.80
	β_1	2.06 (0.48)	2.00 (0.29)	2.00
	β_2	0.97 (0.30)	0.97 (0.15)	1.00
Q-func	β_3	1.16 (0.46)	1.14 (0.29)	1.20
	β_4	2.02 (0.30)	1.97 (0.22)	2.00
	β_5	-0.79 (0.07)	-0.80 (0.03)	-0.80
	β_1	1.95 (0.45)	1.96 (0.33)	2.00
	β_2	0.98 (0.29)	0.96 (0.14)	1.00
mice	β_3	1.17 (0.45)	1.14 (0.32)	1.20
	β_4	1.99 (0.35)	1.97 (0.22)	2.00
	β_5	-0.79 (0.06)	-0.80 (0.03)	-0.80

proposed in Garcia et al. (2010), and select approach based on MICE imputed full data when sample size is 600. When sample size is n=300, proportions of identifying the true model are 61%, 55%, and 48%, respectively for the three methods described. Over-fitting effect is obvious in this case as well, as can be seen from the 0.47, 0.44, and 0.62 zero coefficients being mis-classified as non-zeros (column "IC") in the three algorithms, respectively. Our proposed penalized likelihood approach has the smallest model error, followed by imputation algorithm by FCS and the iterative penalized EM algorithm. The proposed approach on average correctly identifies 4.97 non-zero coefficients and incorrectly identifies 0.07 zero coefficients as non-zeros with the sample size n=600. As sample size increases, all three model selection approaches significantly improve their performance.

Table XIV lists the means of all the non-zero regression coefficients. All three methods yield very good estimates with reasonably good standard deviations.

CHAPTER 7

ANALYSIS OF THE DATA EXAMPLE

In this chapter, we apply the proposed model selection algorithm for logistic regression models with missing values in continuous and binary covariates to the hip fracture data described in chapter 1. We assume that the missing data are MAR. Since the data involve multiple continuous and categorical missing covariates, Monte Carlo simulation along with rejection sampling is used in the E-step of EM algorithm. Once the unpenalized maximum likelihood estimator is obtained, the penalized observed likelihood function with SCAD penalty function can be constructed as in Equation 4.15 and its least square approximation can be obtained. Bayesian Information Criterion is used for selection of the penalty tuning parameter. Since the BIC criteria in the presence of missing data involves observed data log-likelihood as well, the observed data likelihood is approximated by the quadratic expansion, which in turn is approximated by the sum of squares of the first derivatives of Q-function over all observations described in 4.5.

We will compare performance of our proposed method with the model selection method based on imputed full data with missing values imputed by the fully conditional specification approach (Burren et al (2006)). Estimated coefficients will be presented. In addition, we will explore possible interactions of covariates in the variable selection.

The hip fracture data has a total n=436 observations with 17 binary covariates and 10 continuous covariates, along with the outcome variable indicating the presence or absence of hip fracture among male veterans. Since bmi is a perfect function of wt (weight) and ht (height), we remove these two variables. Except the outcome variable and two matching variable race and age are completely observed, all other 23 covariates are subject to missing. We first remove

23 observations that have more than 10 binary variables missing. Among them, 8 observations have missing values in all binary covariates. Then, we standardize the continuous covariates so that it has mean **0** and variance **1**.

The reduced hip fracture data consists of p=25 covariates of which 17 are binary and 8 are continuous. The consecutive regression models have a total of $\frac{p(p+1)}{2} = 325$ non-intercept coefficients. Applying the proposed SCAD-type penalized likelihood model selection method with BIC criterion described in 4.5, with selected tuning parameter $\lambda = 0.16$, the algorithm yields a logistic regression for outcome variable as

$$\begin{aligned} logit(P(y=1|\mathbf{x})) &= -0.68 + 1.28 * x_{etoh} + 1.98 * x_{dementia} \\ &\quad + 2.34 * x_{Antiseiz} - 2.53 * x_{Antichol}. \end{aligned} \tag{7.1}$$

Recall that the complete-case analysis selects 15 non-zeros coefficients. The model fit for the selected model is summarized in Table XV, where a consecutive regression model is applied to each of the variable in a format that each one is regressed on the rest of the covariates. Regressors with non-zero coefficients are listed in the "Regressors" columns from which it can be seen that at most 5 regressors being selected for each of the covariates' model. From the table, it can be seen that in total 27 coefficients out of 325 are selected to be non-zeros.

To fully utilize the original data, we apply Burren's imputation method to those 23 observations with more than 10 missing values in binary covariates. Since time for calculating conditional expectations in Q-function increases significantly when missing values in binary variable is more than 10, we imputed these observations to reduce their missing values in binary variables so that the number of missing values in binary variables are less than 10. Then we apply the partially imputed data to our proposed method of model selection in maximizing

TABLE XV

SELECTED COVARIATE REGRESSION MODELS BY PROPOSED LIKELIHOOD METHOD WITH N=413

	Regressors				
Variable	1	2	3	4	5
etoh	smoke	-	-	-	-
smoke	Steroids	COPD	-	-	-
CVA	seizure	LevoT4	-	-	-
dementia	seizure	Sedat	-	-	-
Parkinson	Antiseiz	bmi	-	-	-
seizure	Sedat	Antiseiz	CaCO3	LevoT4	-
sedat	HCTZ	Antiseiz	LevoT4	-	-
NSAIDS	-	-	-	-	-
Steroids	Lasix	CaCo3	LevoT4	-	-
Lasix	LevoT4	-	-	-	-
HCTZ	-	-	-	-	-
Antiseiz	COPD	-	-	-	-
CaCo3	-	-	-	-	-
LevoT4	Antichol	-	-	-	-
Antichol	-	-	-	-	-
COPD	-	-	-	-	-
age	-	-	-	-	-
bmi	-	-	-	-	-
hgb	hct	-	-	-	-
hct	-	-	-	-	-
BUN	-	-	-	-	-
Cr	-	-	-	-	-
albumin	-	-	-	-	-
cholesterol	-	-	-	-	-
age	-	-	-	-	-
race	-	-	-	-	-

the SCAD-type penalized likelihood described in 4.4 with BIC selection of tuning parameter described in 4.5. The fitted logistic regression model for outcome variable is

$$\begin{aligned} logit(P(y=1|\mathbf{x})) &= -0.48 + 1.03 * x_{etoh} + 1.92 * x_{dementia} \\ &\quad + 1.33 * x_{Antiseiz} - 1.98 * x_{Antichol}. \end{aligned} \tag{7.2}$$

We see that the same covariates are selected as Equation 7.1.

The fitted covariates' model is summarized in Table XVI with similar format as Table XV. For this data set, 36 coefficients are selected to be non-zeros in 325 of them.

We see the model selection algorithm greatly reduces the model dimension and gives a much simpler dependence structure for the covariates. For comparison, we first apply Burren's method (Burren et al., 2006) to impute the missing values and apply the penalized likelihood method to the imputed full data to select variables. To remove the uncertainty of imputation, we repeat the imputation-selection process for 100 times and summarize the frequencies of the regression model for outcome variable in Table XVII.

From Table XVII, we can see that only variables *etoh*, *dementia*, *Antiseiz* and *AntiChol* have more than 50% frequencies to be selected. It confirms with previous results obtained from proposed model selection in maximizing penalized likelihood. By averaging its regression coefficients, the resulting logistic regression model for the outcome is

$$\begin{aligned} logit(P(y=1|\mathbf{x})) &= -0.53 + 0.98 * x_{etoh} + 1.60 * x_{dementia} \\ &\quad + 1.82 * x_{Antiseiz} - 1.35 * x_{Antichol}. \end{aligned} \tag{7.3}$$

TABLE XVI

SELECTED COVARIATE REGRESSION MODELS BY PROPOSED LIKELIHOOD METHOD WITH N=436

	Regressors					
Variable	1	2	3	4	5	6
etoh	smoke	-	-	-	-	-
smoke	Parkinson	Steroids	COPD	-	-	-
CVA	seizure	LevoT4	-	-	-	-
dementia	seizure	Sedat	-	-	-	-
Parkinson	seizure	Lasix	CaCO3	LevoT4	COPD	bmi
seizure	Sedat	HCTZ	Antiseiz	LevoT4	-	-
sedat	HCTZ	Antiseiz	LevoT4	-	-	-
NSAIDS	-	-	-	-	-	-
Steroids	Lasix	CaCo3	LevoT4	COPD	-	-
Lasix	HCTZ	LevoT4	-	-	-	-
HCTZ	-	-	-	-	-	-
Antiseiz	COPD	-	-	-	-	-
CaCo3	hgb	hct	-	-	-	-
LevoT4	Antichol	-	-	-	-	-
Antichol	-	-	-	-	-	-
COPD	-	-	-	-	-	-
bmi	-	-	-	-	-	-
hgb	hct	-	-	-	-	-
hct	-	-	-	-	-	-
BUN	-	-	-	-	-	-
Cr	-	-	-	-	-	-
albumin	-	-	-	-	-	-
cholesterol	-	-	-	-	-	-
age	-	-	-	-	-	-
race	-	-	-	-	-	-

TABLE XVII

COVARIATE SELECTED FREQUENCY FOR THE OUTCOME REGRESSION MODEL BY BURREN'S IMPUTATION METHOD (BURREN ET AL., (2006))

Covariate	Frequency
etoh	0.94
smoke	0.22
CVA	0.00
dementia	0.99
parkinson	0.00
seizure	0.00
Sedat	0.00
NSAIDs	0.00
Steroids	0.01
Lasix	0.02
HCTZ	0.00
Antiseiz	0.99
CaCO3	0.00
LevoT4	0.42
AntiChol	0.86
COPD	0.00
bmi	0.00
hgb	0.21
hct	0.03
BUN	0.00
Cr	0.00
albumin	0.00
cholesterol	0.00
age	0.00
race	0.00

To check the stability of the proposed method in model selection, we permute the original order of the covariates and re-ran the analysis. By reversing the order within binary variables and continuous variables, the fitted SCAD sparse estimate gives the following regression model

$$\begin{aligned} logit(P(y=1|\mathbf{x})) &= -0.54 + 1.01 * x_{etoh} + 1.21 * x_{dementia} \\ &\quad + 1.51 * x_{Antiseiz} - 1.18 * x_{AntiChol}. \end{aligned} \tag{7.4}$$

This model also indicates that etoh, dementia, Antiseiz, and Antichol, which are included in the previous three selection results, are significant.

From the model selection results given in Equation 7.1, Equation 7.2, Equation 7.3, and Equation 7.4, we can see that only four covariates are significant in predicting the presence of hip fracture: etoh, dementia, Antiseiz, and Antichol. Three other covariates, smoke, LevoT4, hgb, are possibly selected. Next, we include these four significant variables, smoke, LevoT4, hgb, and their iterations into the model. In order to maintain the validity of the covariates models, we only consider those interaction terms for the regression model of the outcome variable. We use the imputed data set by FCS method without interaction then perform model selection with interactions. Repeat this process for each interaction term for 100 times and we check the frequencies for any selected interaction. We summarize results of this process in Table XVIII. From the proportion of interaction terms are being selected, we know that three interaction terms have significant effects for the presence of hip fracture: etoh-dementia, dementia-smoke, and Antiseiz-smoke. The other possible significant interaction terms are etoh-Antiseiz, etoh-AntiChol, and AntiChol-smoke. Last analysis, we use the imputed data set by FCS and put all potential 6 interaction terms into the data for model selection. The frequencies of being selected are summarized in Table XIX. From Table XIX, interaction terms

TABLE XVIII

INTERACTION SELECTED FREQUENCIES FOR THE OUTCOME REGRESSION MODEL BY BURREN'S IMPUTATION METHOD (BURREN ET AL., (2006))

Interaction Terms	Frequency
etoh-dementia	0.79
etoh-Antiseiz	0.31
etoh-AntiChol	0.47
etoh-smoke	0.07
etoh-LevoT4	0.00
etoh-hgb	0.00
dementia-Antiseiz	0.05
dementia-AntiChol	0.00
dementia-smoke	0.87
dementia-LevoT4	0.00
dementia-hgb	0.00
Antiseiz-AntiChol	0.00
Antiseiz-smoke	0.85
Antiseiz-LevoT4	0.00
Antiseiz-hgb	0.00
AntiChol-smoke	0.34
AntiChol-LevoT4	0.40
AntiChol-hgb	0.00
smoke-LevoT4	0.00
smoke-hgb	0.00
LevoT4-hgb	0.00

etoh-dementia, dementia-smoke, and Antiseiz-smoke are confirmed to have an significant effect on the outcome. Including them with the four significant variables identified earlier yields the following one-step fit

$$\begin{aligned} & logit(P(y=1|\mathbf{x})) \\ = \; & -0.97 + 0.96 * x_{etoh} + 0.79 * x_{smoke} + 1.98 * x_{dementia} + 1.67 * x_{Antiseiz} - 1.52 * x_{AntiChol} \\ & -1.86 * x_{etoh-dementia} + 1.87 * x_{dementia-smoke} + 1.59 * x_{Antiseiz-smoke}. \end{aligned} \tag{7.5}$$

TABLE XIX

SIX INTERACTION TERMS SELECTED FREQUENCIES FOR THE OUTCOME REGRESSION MODEL BY BURREN'S IMPUTATION METHOD (BURREN ET AL., (2006))

Interaction Terms	Frequency
etoh-dementia	0.96
etoh-Antiseiz	0.28
etoh-AntiChol	0.48
dementia-smoke	0.96
Antiseiz-smoke	0.86
AntiChol-smoke	0.45

CHAPTER 8

CONCLUSION

The problem of variable selection is common in regression models. The presence of missing values in data increases the difficulty of applying model selection methods of maximizing penalized likelihood with LASSO or SCAD penalty. The method of maximizing SCAD type penalized likelihood was proposed in the previous work by Fan and Li (2001). It was showed that it can give consistent sparse estimator of the coefficients for the underlying true model. By iteratively maximizing the penalized Q-function, Garcia et al. (2010), applied the model selection to Cox regression models with missing data with an IC_Q criteria suggested in Ibrahim et al. (2008). However, the model selection algorithm mentioned above did not utilize the observed data log-likelihood so that the MPLE may not have the good properties obtained in Fan and Li (2001) on model selection. Therefore, how to apply the maximizing penalized likelihood method to generalized linear models, including multiple linear regression and logistic regression, remains a challenge for methodology development. For such a purpose, in this dissertation, we proposed a new method of maximizing penalized observed data log-likelihood with SCAD penalty in linear regression models with missing data and extend it to logistic regression models.

We propose not only select covariates for the outcome regression model, but also extend traditional model selection process to the whole data structure, allowing model selections within covariate models. By doing so, scientific researchers will have a better understanding in covariates dependency relationships. For variable selection with missing covariates in high-dimension data with $p \gg n$, only applying model selection for the primary outcome variable may be problematic because of the high dimensionality of covariate models. However, extending

model selection to covariates' models in the proposed method can potentially help tackle this problem.

We proposed to use the BIC criterion (Schwarz, 1978) for selecting optimal tuning parameter in penalized likelihood with SCAD penalty. For implementing our proposed algorithm in logistic regression models, three R programs are developed to fit the model with arbitrary missing patterns in continuous, binary and mixed continuous and binary covariates, respectively.

We perform simulation studies in linear regression with missing covariates to demonstrate the performance of the proposed algorithm, comparing with the method proposed in Ibrahim et al. (2008), and Garcia et al. (2010). We create the mean and covariance structure such that the true coefficients in the outcome and covariates regression model have a sparse format. The MRME of the selected model to that of the unpenalized estimators indicates that the sparse estimates from both algorithms dramatically reduce model error and the ability of correctly identifying the truth is increased when sample size increases. Our proposed algorithm of maximizing the penalized observed likelihood with BIC criterion outperforms the one proposed by Garcia et al. (2010), with ICQ suggested in Ibrahim et al. (2008), in identifying more correct-fit in model selection process.

We extend our proposed algorithm to logistic regression models with missing covariates, in which the expected full data log-likelihood conditional on the observed data does not have a closed form. Gauss-Hermite Quadrature is used to compute the intractable integrations in the EM algorithm when at most two covariates are subject to missing. When more continuous covariates are subject to missing, Gauss-Hermite Quadrature becomes less and less inefficient in computations. We therefore use Monte Carlo simulation along with rejection sampling in computing the Q-function in the EM algorithm. We run simulation with 200 replications of data

sets with missing covariates with different sample size and compare the performance with the method proposed in Garcia et al. (2010), and the imputation-selection method based on FCS by Burren et al. (2006). Our simulation shows that our proposed algorithm outperforms the other alternative methods with respect to proportion of correct-fit of the data with moderate sample size (n=600).

In the section of data analysis, we apply the proposed model selection algorithm to the hip fracture data to illustrate the application of the method. The data is from a case-control study to investigate potential risk factors of hip fracture among male veterans. Since the data have many covariates missing patterns, we reshape the data by removing some observations with more than 10 missingness in binary covariates and apply the proposed model selection algorithm to it. We also use Burren's FCS imputation algorithm to partially impute those observations so that we can apply our algorithm to the full data with 436 observations. The model selection results for these two data sets show that only four covariates out of 27 are significant predictors of the outcome. They give similar coefficient values but the first case with less subjects gives larger estimated coefficients. We also implement the model selection by repeatedly imputed data set using Burren's FCS method and apply our proposed model selection method to the permuted data to validate our conclusion about important covariates. Last, we consider more potential risk factors and their interaction effects in the regression model for outcome variable based on the imputed data set.

CITED LITERATURE

1. Abramowitz, M. and Stegun, I. A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover Publications, New York, 1972.

2. Adams, J.: A computer experiment to evaluate regression strategies. Journal of the American Statistical Association, In Proceedings of the Statistical Computing Section:55–62, 1990.

3. Barengolts, E., Karanouh, D., Kolodny, L., and Kukreja, S.: Risk factors for hip fractures in predominantly african-american veteran male population. Journal of Bone and Mineral Research, 16(S170), 2001.

4. Breiman, L.: Better subset regression using the nonnegative garrote. Technometrics, 37(4):373–384, 1995.

5. Breiman, L.: Heuristics of instability and stabilization in model selection. The Annals of Statistics, 24(6):2350–2383, 1996.

6. Breslow, N. E.: Analysis of survival data under proportional hazards model. Intemational Statistical Review, 43(1):45–58, 1975.

7. Cai, Z., Fan, J., and Li, R.: Efficient estimation and inferences for varying-coefficient models. Journal of the American Statistical Association, 95(451):888–902, 2000.

8. Casella, G. and George, E. I.: Explaining the gibbs sampler. The American Statistician, 46(3):167–174, 1992.

9. Cox, D.: Regression models and life tables (with discussion). Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–220, 1972.

10. Dempster, A., Laird, N., and Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38, 1977.

11. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R.: Least angle regression. The Annals of Statistics, 32(2):407–499, 2004.

12. Efroymson, M.: Multiple regression analysis, Mathematical Methods for Digital Computers (Ralston, A. and Wilf, H.S., ed.). Wiley, New York, 1960.

13. Evans, M. and Swartz, T.: Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press, Oxford, 2000.

14. Fan, J. and Chen, J.: One-step local quasi-likelihood estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(4):927–943, 1999.

15. Fan, J. and Gijbels, I.: Local Polynomial Modeling and Its Applications. Chapman and Hall, London, 1996.

16. Fan, J. and Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360, 2001.

17. Frank, I. and Friedman, J.: A statistical view of some chemometrics regression tools. Technometrics, 35(2):109–135, 1993.

18. Fu, W.: Penalized regressions: The bridge versus the lasso. Journal of Computational and Graphical Statistics, 7(3):397–416, 1998.

19. Garcia, R., Ibrahim, J., and Zhu, H.: Variable selection in the cox regression model with covariates missing at random. Biometrics, 66(1):97–104, 2010.

20. Gelfand, A. E. and Smith, A. F. M.: Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85(410):398–409, 1990.

21. Geman, S. and Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions of Pattern Analysis and Machine Intelligence, 6:721–741, 1984.

22. George, E. I. and McCulloch, R. E.: Variable selection via gibbs sampling. Journal of the American Statistical Association, 88(423):881–889, 1993.

23. Gilks, W. R. and Wild, P.: Adaptive rejection sampling for gibbs sampling. Journal of the Royal Statistical Society. Series C (Applied Statistics), 41(2):337–348, 1992.

24. Givens, G. and Hoeting, J.: Computational Statistics. Wiley, New York, 2005.

25. Herring, A. H. and Ibrahim, J. G.: Likelihood-based methods for missing covariates in the cox proportional hazards model. Journal of the American Statistical Association, 96(453):292–302, 2001.

26. Hoerl, A. and Kennard, R.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970a.

27. Hoerl, A. and Kennard, R.: Ridge regression: applications to nonorthogonal problems. Technometrics, 12(1):69–82, 1970b.

28. Hoerl, R., Schuenemeyer, J., and Hoerl, A.: A simulation of biased estimation and subset selection regression technique. Technometrics, 28(4):369–380, 1986.

29. Hurvich, C. and Tsai, C.-L.: The impact of model selection on inference in linear regression. The American Statistician, 44(3):214–217, 1990.

30. Ibrahim, J. G., Lipsitz, S. R., and Chen, M. H.: Missing covariates in generalized linear models when the missing data mechanism is nonignorable. Journal of the Royal Statistical Society: Series B, 61(1):173–190, 1999.

31. Ibrahim, J., Zhu, H., and Tang, N.: Model selection criteria for missing-data problems using the em algorithm. Journal of the American Statistical Association, 103(484):1648–1658, 2008.

32. Knight, K. and Fu, W.: Asymptotics for lasso-type estimators. The Annals of Statistics, 28(5):1356–1378, 2000.

33. Little, R. and Rubin, D.: Statistical Analysis with Missing Data, 2nd edition. John Wiley, New York, 2002.

34. Louis, T. A.: Finding the observed information matrix when using the em algorithm. Journal of the Royal Statistical Society: Series B, 44(2):226–233, 1982.

35. Meinshausen, N. and Buhlmann, P.: Variable selection and high-dimensional graphs with the lasso. The Annals of Statistics, 34(3):1436–1462, 2004.

36. Miller, A.: Subset Selection in Regression. Chapman and Hall, London, 1990.

37. Nelder, J. and Wedderburn, R.: Generalized linear models. Journal of the Royal Statistical Society: Series A, 135(3):370–384, 1972.

38. Osborne, M., Presnell, B., and Turlach, B.: On the lasso and its dual. Journal of Computational and Graphical Statistics, 9(2):319–337, 2000.

39. Press, W., Teukolsky, S., Vetterling, W., and Flannery, B.: Numerical Recipes: The Art of Scientific Computing (3rd ed.). Cambridge University Press, New York, 2007.

40. Robins, J. M., Rotnitzky, A., and Zhao, L. P.: Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89(427):846–866, 1994.

41. Robins JM, R. and Fei, Z. L.: Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, (89):846–866, 1994.

42. Roecker, E.: Prediction error and its estimation for subset-selected models. Technometrics, 33(4):459–468, 1991.

43. Rubin, D.: Inference and missing data. Biometrika, 63(3):581–592, 1976.

44. Rubin, D.: Multiple imputation for nonresponse in surveys. John Wiley, New York, 1987.

45. Rubin, D.: Multiple imputation after 18+ years. Journal of the American Statistical Association, 91(434):473–489, 1996.

46. Schafer, J. L.: Analysis of Incomplete Multivariate Data. Chapman and Hall, London, 1997.

47. Schafer, J. L.: Multiple imputation: A primer. Statistical Methods in Medical Research, 8(1):3–15, 1999.

48. Schwarz, G. E.: Estimating the dimension of a model. Annals of Statistics, 6(2):461–464, 1978.

49. Stroud, A. H. and Secrest, D.: Gaussian Quadrature Formulas. Prentice-Hall series in automatic computation. Prentice-Hall, Englewood Cliffs (N.J.), 1966.

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

51. Wang, H. and Leng, C.: Unified lasso estimation by least squares approximation. Journal of the American Statistical Association, 102(479):1039–1048, 2007.

52. Wang, H., Li, R., and Tsai, C.-L.: Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, 94(3):553–568, 2007.

53. Wei, G. C. G. and Tanner, M. A.: A monte carlo implementation of the em algorithm and the poor mans data augmentation algorithms. Journal of the American Statistical Association, 85(411):699–704, 1990.

54. Yang, X., Belin, T., and Boscardin, W.: Imputation and variable selection in linear regression models with missing covariates. Biometrics, 61(2):498–506, 2005.

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

56. Zou, H. and Li, R.: One-step sparse estimates in nonconcave penalized likelihood models. The Annals of Statistics, 36(4):1509–1533, 2008.

VITA

NAME:	Fei Shi
EDUCATION:	B.S., Scientific Computation and Computer Applications, Sun Yat-Sen University, China, 2004. M.S., Applied Mathematics, University of Illinois at Urbana-Champaign, Urbana, Illinois, 2005. Ph.D., Biostatistics, University of Illinois at Chicago, Chicago, Illinois, 2012.
TEACHING EXPERIENCE:	Department of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago: Power and Sample Size Calculation in Design of Clinical Trials for Graduates, 2012.
RESEARCH EXPERIENCE:	Research Assistant, Department of Pathology, College of Medicine, University of Illinois at Chicago, 2011-2012. Research Specialist, Center for Molecular Biology of Oral Diseases, College of Dentistry, University of Illinois at Chicago, 2008-2010.
PROFESSIONAL EXPERIENCE:	Clinical Statistician, Epidemiology Surveillance & Pharmacoepidemiology, Abbott Laboratories, Libertyville, Illinois, 2010-2011 Statistical Analyst Intern, Discover Financial Services, Riverwoods, Illinois, May-August, 2007.

PUBLICATIONS:

Liu, X., Wang, A., Lo Muzio, L., Kolokythas, A., Sheng, S., Rubini, C., Ye, H., Shi, F., Yu, T., Crowe, D., Zhou, X.. Deregulation of manganese superoxide dismutase (SOD2) expression and lymph node metastasis in tongue squamous cell carcinoma. *BMC Cancer*, 2010, 10:365.

Jiang, L., Liu X., Kolokythas, A., Yu, J., Wang A., Heidbreder, C.E., Shi, F.; Zhou, X.. Downregulation of the Rho GTPase signaling pathway is involved in the microRNA-138-mediated inhibition of cell migration and invasion in tongue squamous cell carcinoma. *Int J Cancer.*, 2010 Aug 1, 127(3):505-12.

Demirtas H., A. Amatya, O. Pugach, J. F. Cursio, F. Shi, D. Morton, and B. Doganay. Accuracy versus convenience: a simulation-based comparison of two continuous imputation models for incomplete ordinal longitudinal clinical trials data. *Statistics and Its Interface*, 2009, 2(4), 449-456.

Wang A., Liu X., Sheng S., Ye H., Peng T., Shi, F., Crowe D.L., Zhou X. (2009). Dysregulation of heat shock protein 27 expression in tongue squamous cell carcinoma.*BMC Cancer*, 2009 Jun 4, 9:167.

Liu X, Yu J, Jiang L, Wang A, Shi, F., Ye H, Zhou X. MicroRNA-222 regulates cell invasion by targeting matrix metalloproteinase 1 (MMP1) and manganese superoxide dismutase 2 (SOD2) in tongue squamous cell carcinoma cell lines. *Cancer Genomics Proteomics*, 2009 May-Jun; 6(3):131-139.

Liu X, Jiang L, Wang A, Yu J, Shi, F., Zhou X., MicroRNA-138 inhibits invasion in oral squamous cell carcinoma cell lines. *Cancer Lett*, 2009 Dec 28; 286(2):217-22. Epub 2009 Jun 21.