Statistica Sinica **20** (2010), 149-165

VARIABLE SELECTION FOR REGRESSION MODELS WITH MISSING DATA

Ramon I. Garcia, Joseph G. Ibrahim and Hongtu Zhu

University of North Carolina at Chapel Hill

Abstract: We consider the variable selection problem for a class of statistical models with missing data, including missing covariate and/or response data. We investigate the smoothly clipped absolute deviation penalty (SCAD) and adaptive LASSO and propose a unified model selection and estimation procedure for use in the presence of missing data. We develop a computationally attractive algorithm for simultaneously optimizing the penalized likelihood function and estimating the penalty parameters. Particularly, we propose to use a model selection criterion, called the IC_Q statistic, for selecting the penalty parameters. We show that the variable selection procedure based on IC_Q automatically and consistently selects the important covariates and leads to efficient estimates with oracle properties. The methodology is very general and can be applied to numerous situations involving missing data, from covariates missing at random in arbitrary regression models to nonignorably missing longitudinal responses and/or covariates. Simulations are given to demonstrate the methodology and examine the finite sample performance of the variable selection procedures. Melanoma data from a cancer clinical trial is presented to illustrate the proposed methodology.

Key words and phrases: EM algorithm, IC_Q, missing data, penalized likelihood, variable selection.

1. Introduction

Variable selection procedures based on penalized likelihood methods have received much attention in the recent literature (Bickel and Li (2006)). Some notable methods include the Lasso, Smoothly Clipped Absolute Deviation penalty (SCAD) (Fan and Li (2001)), and Adaptive Lasso (ALASSO) (Zou (2006)), among many others. These methods have been successfully applied to generalized linear models and robust linear regression (Fan and Li (2001)), and to semiparametric models including Cox's proportional hazards model (Fan and Li (2002, 2004)). Moreover, under an appropriate choice of the penalty parameter, these variable selection procedures can produce efficient estimates with oracle properties (Fan and Li (2001)). The methods for selecting the penalty parameters consist of minimizing the penalty parameter with respect to some

criterion. Commonly used criteria include generalized cross-validation (GCV) and the Bayesian Information Criterion (BIC). It has been shown that BIC can identify the true model consistently, whereas GCV cannot (Wang, Li and Tsai (2007)). Ideally, one would like to use a criterion that results in appropriate choices of the penalty parameter so that the penalized likelihood estimates can possess oracle properties. However, to the best of our knowledge, a general and easy-to-compute penalty and variable selection procedure is not currently available for missing data problems.

Missing data are a common problem in various settings, including surveys, clinical trials, and longitudinal studies. Responses and/or covariates may be missing, and statistical models for handling the missing data often depend on the missing data mechanism, such as data not missing at random (NMAR), also referred to as nonignorable missingness. For example, when there are NMAR covariates, one must specify both the covariate distribution and the missing data mechanism in the likelihood function. These additional distributions bring additional parameters into the model, that need to be taken into consideration in model selection. It is common to use some model selection criterion, such as AIC and BIC, based on the observed data log-likelihood to select a small set of variables. For instance, one might use AIC (or BIC) to select a small subset of 'covariates' that best predicts the outcome of interest. However, even in the absence of missing data, model selection criteria, such as AIC, can become infeasible for variable selection in linear regression models with a large number of covariates (Fan and Li (2001, 2002)). More discussion on the drawbacks of best subset selection can be found in Fan and Li (2001).

Performing variable selection in statistical models for missing data problems raises several new statistical challenges, underscoring the need for methodological development. In many missing data problems, the observed data log-likelihood does not have a closed form and is often computationally intractable because it requires evaluation of high dimensional integrals which do not have a closed form. These integrals can be approximated but the accuracy of the approximation is essentially impossible to assess in many cases. Thus, it can be infeasible to directly maximize the observed data log-likelihood function, along with the SCAD or ALASSO penalties, to select important variables and calculate their estimates. Furthermore, computing the GCV and BIC to select the penalty parameter also requires computing the intractable likelihood function and running an optimization algorithm for each penalty parameter, which can be computationally intensive for missing data problems. Thus, it is also critical to develop a new penalty selection criterion, that is easy-to-compute, in missing data problems.

The aim of this paper is to develop variable selection and penalty selection procedures, along with the SCAD and ALASSO penalties, for a class of statistical models in missing data problems, including generalized linear models with missing covariates and/or responses, random effects models, and latent variable models. We reformulate the penalty parameters in the SCAD and ALASSO as a hyperparameter in the model, and then we use the EM algorithm to simultaneously optimize the penalized likelihood function and estimate the penalty parameters. In addition, we also develop an alternative method based on optimizing a new criterion, which we call the IC_Q criterion, to select penalty parameters. The variable selection and penalty selection procedures developed here are very general and can be applied to numerous situations involving missing data and/or random effects and latent variables. Under some regularity conditions, we establish the asymptotic properties (e.g., oracle properties) of the penalized maximum likelihood estimator and the consistency of the IC_Q-based penalty selection procedure.

The rest of the paper is organized as follows. Section 2 gives the general development of algorithms for maximizing the penalized likelihood function and selecting penalty parameters in missing data problems; we characterize the asymptotic properties of the penalized maximum likelihood (ML) estimator and the IC_Q penalty selection procedure. Section 4 first presents a simulation study involving missing at random (MAR) covariates in linear models in order to examine the finite sample performance of the penalized ML estimates using various penalty parameter selection procedures. In Section 4, a Melanoma dataset is analyzed with the proposed methodology. We conclude the paper with some discussion in Section 5.

2. Variable Selection for Regression Models with Missing Data

2.1. Model formulation

For notational simplicity, we focus on data with MAR or NMAR covariates; however, the methods developed below can be adapted to data with both missing responses and covariates (see Ibrahim, Lipsitz and Chen (2001)). Suppose there are n independent observations $(\mathbf{x}_1, \mathbf{z}_1, y_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n, y_n)$, where y_i is the response variable, $\mathbf{z}_i$ is a $q \times 1$ vector of partially observed covariates, and $\mathbf{x}_i$ is a $(p-q)\times 1$ vector of completely observed covariates. Let $\mathbf{z}_{m,i}$ and $\mathbf{z}_{o,i}$, respectively, denote the missing and observed components of $\mathbf{z}_i$. We use the $q \times 1$ random vector $\mathbf{r}_i$ to indicate the missingness of $\mathbf{z}_i$, where the k^{th} component $r_{ik} = 1$ when z_{ik} is observed and $r_{ik} = 0$ when z_{ik} is missing. We denote the complete and

observed data of subject i by $\mathbf{D}_{c,i}$ and $\mathbf{D}_{o,i}$, respectively, and the entire complete and observed data by $\mathbf{D}_c$ and $\mathbf{D}_o$, respectively.

When the covariates are NMAR, the complete data likelihood is the product of the joint distribution of $(y_i, \mathbf{z}_i, \mathbf{r}_i)$ given $\mathbf{x}_i$, denoted by $f(y_i, \mathbf{z}_i, \mathbf{r}_i|\mathbf{x}_i)$, which is typically specified as a product of three conditional distributions as

$$f(\mathbf{D}_c) = \prod_{i=1}^{n} f(y_i, \mathbf{z}_i, \mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\eta}) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\tau}) f(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\alpha}) f(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\xi}), \tag{2.1}$$

where $\boldsymbol{\eta} = (\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\xi})$ are the parameters corresponding to response model, covariate distribution, and missing data mechanism. We use the generic label $f(u_1|u_2)$ throughout to denote the conditional distribution of u_1 given u_2. If the covariates are MAR, then the missing data mechanism, $f(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\xi})$, can be ignored from (2.1).

As in generalized linear models (see McCullagh and Nelder (1989, Chap.2)), we assume that the conditional distribution of y_i given $(\mathbf{x}_i, \mathbf{z}_i)$, denoted by $f(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta, \boldsymbol{\tau})$, satisfies

$$E[y_i|\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}, \boldsymbol{\tau}] = \mu_i = g\Big((\mathbf{x}_i^T, \mathbf{z}_i^T)\boldsymbol{\beta}\Big), \tag{2.2}$$

where $\boldsymbol{\tau}$ denotes the additional parameters in $f(y_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\tau})$, $g(\cdot)$ is a known link function, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of regression coefficients. In practice, it is common to assume that y_i given $(\mathbf{x}_i, \mathbf{z}_i)$ belongs to the exponential family, such as the binomial, normal, Poisson, etc.. (Little and Schluchter (1985), and Ibrahim and Lipsitz (1996)).

We model the missing-data mechanism for NMAR covariates according to either a joint log-linear model for $f(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\xi})$ or a product of a sequence of one dimensional conditionals as in Ibrahim, Chen and Lipsitz (1999). Finally, we assume that the covariate distribution $f(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\alpha})$ is also modeled via a sequence of one-dimensional conditional distributions as in and Ibrahim, Chen and Lipsitz (1999), and is given by

$$f(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\alpha}) = f(z_{iq}|z_{i(q-1)}, \ldots, z_{i1}, \mathbf{x}_i, \boldsymbol{\alpha}) \times \cdots f(z_{i1}|\mathbf{x}_i, \boldsymbol{\alpha}),$$

where we assume a specific order of conditioning.

2.2. Penalized likelihood for variable selection

In the variable selection problem, our objective is to identify nonzero components of $\boldsymbol{\beta}$ in (2.2) and simultaneously estimate parameters, while accounting

for the missing covariate data. We propose to maximize the penalized likelihood function given by

$$P(\boldsymbol{\eta}|\boldsymbol{\lambda}) = \sum_{i=1}^{n} \log f(\mathbf{D}_{o,i}|\boldsymbol{\eta}) - n\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|), \tag{2.3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$, λ_j is the penalty parameter corresponding to the j-th regression coefficient β_j, and $f(\mathbf{D}_{o,i}|\boldsymbol{\eta}) = \int f(y_i, \mathbf{z}_i, \mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\eta}) d\mathbf{z}_{m,i}$ is the observed-data log-likelihood function of the i-th observation. The penalty function, $p_{\lambda_j}(\cdot)$, is a nonnegative, nondecreasing, and differentiable function on $(0, \infty)$ (Fan and Li (2001) and Zou (2006)). These properties ensure that the maximization of (2.3) results in estimates of $\boldsymbol{\beta}$ which are shrunk to zero if they are small. The corresponding covariates of the estimates that are zero are the insignificant predictors of the response variable, whereas the estimates that are not zero correspond to those covariates which are statistically significant predictors. By maximizing (2.3), one can select significant predictors and estimate parameters simultaneously while accounting for the missing data. This approach is in sharp contrast to stepwise selection procedures and Bayesian procedures (George and McCulloch (1993), and Yang, Belin and Boscardin (2005)), that ignore stochastic errors inherited in the selection phase during estimation of the 'best' model (Fan and Li (2002)).

In (2.3), the parameters $\boldsymbol{\tau}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\xi}$ are not penalized, so they are not shrunk to zero even though their actual values may be small. In this sense, variable selection does not occur in the covariate distribution and the missing data mechanism. However, care must be taken in the specification of these distributions since certain specifications can lead to identifiability issues for estimating $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$, and thus $\boldsymbol{\beta}$.

Because the observed-data log-likelihood function usually involves intractable integration, we use the EM algorithm to compute the penalized maximum likelihood estimate of $\boldsymbol{\eta}$, denoted by $\widehat{\boldsymbol{\eta}}_\lambda$, for each $\boldsymbol{\lambda}$ (Dempster, Laird and Rubin (1977)). At the s-th iteration, given $\boldsymbol{\eta}^{(s)}$, the E step is to evaluate the $Q-$function given by

$$\begin{aligned}
Q_\lambda(\boldsymbol{\eta}|\boldsymbol{\eta}^{(s)}) &= E\Big[\log f(\mathbf{D}_c|\boldsymbol{\eta})|\mathbf{D}_o, \boldsymbol{\eta}^{(s)}\Big] - n\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \\
&= Q(\boldsymbol{\eta}|\boldsymbol{\eta}^{(s)}) - n\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \\
&= Q_1(\boldsymbol{\beta}, \boldsymbol{\tau}|\boldsymbol{\eta}^{(s)}) - n\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) + Q_2(\boldsymbol{\alpha}|\boldsymbol{\eta}^{(s)}) + Q_3(\boldsymbol{\xi}|\boldsymbol{\eta}^{(s)}) \\
&= Q_{1,\lambda}(\boldsymbol{\beta}, \boldsymbol{\tau}|\boldsymbol{\eta}^{(s)}) + Q_2(\boldsymbol{\alpha}|\boldsymbol{\eta}^{(s)}) + Q_3(\boldsymbol{\xi}|\boldsymbol{\eta}^{(s)}),
\end{aligned}$$

where

$$Q_3(\boldsymbol{\xi}|\boldsymbol{\eta}^{(s)}) = \int \sum_{i=1}^{n} \log\Big[f(\mathbf{r}_i|y_i,\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\xi})\Big] f(\mathbf{z}_{m,i}|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\mathbf{r}_i,\boldsymbol{\eta}^{(s)})d\mathbf{z}_{m,i},$$

$$Q_2(\boldsymbol{\alpha}|\boldsymbol{\eta}^{(s)}) = \int \sum_{i=1}^{n} \log\Big[f(\mathbf{z}_i|\mathbf{x}_i,\boldsymbol{\alpha})\Big] f(\mathbf{z}_{m,i}|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\mathbf{r}_i,\boldsymbol{\eta}^{(s)})d\mathbf{z}_{m,i},$$

$$Q_{1,\lambda}(\boldsymbol{\beta},\boldsymbol{\tau}|\boldsymbol{\eta}^{(s)}) = \int \sum_{i=1}^{n} \log\Big[f(y_i|\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\beta},\boldsymbol{\tau})\Big] f(\mathbf{z}_{m,i}|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\mathbf{r}_i,\boldsymbol{\eta}^{(s)})d\mathbf{z}_{m,i} - n\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|).$$

The M step of the algorithm involves maximizing $Q_{1,\lambda}(\boldsymbol{\beta},\boldsymbol{\tau}|\boldsymbol{\eta}^{(s)})$, $Q_2(\boldsymbol{\alpha}|\boldsymbol{\eta}^{(s)})$, and $Q_3(\boldsymbol{\xi}|\boldsymbol{\eta}^{(s)})$, independently. Maximizing $Q_\lambda(\boldsymbol{\eta}|\boldsymbol{\eta}^{(s)})$ with respect to $(\boldsymbol{\alpha},\boldsymbol{\tau},\boldsymbol{\xi})$ can be done using standard maximization algorithms, such as Newton-Raphson (Little and Schluchter (1985), and Ibrahim and Lipsitz (1996)). However, it is difficult to maximize $Q_{1,\lambda}(\boldsymbol{\beta},\boldsymbol{\tau}^{(s)}|\boldsymbol{\eta}^{(s)})$ with respect to $\boldsymbol{\beta}$, because it is nondifferentiable and nonconcave (Zou and Li (2008)).

To maximize $Q_{1,\lambda}(\boldsymbol{\beta},\boldsymbol{\tau}^{(s)}|\boldsymbol{\eta}^{(s)})$ with respect to $\boldsymbol{\beta}$, we approximate $Q_1(\boldsymbol{\beta},\boldsymbol{\tau}^{(s)}|\boldsymbol{\eta}^{(s)})$ using a second order Taylor's series expansion centered at $\boldsymbol{\beta}^{(s)}$. Using this approximation, $Q_{1,\lambda}(\boldsymbol{\beta},\boldsymbol{\tau}^{(s)}|\boldsymbol{\eta}^{(s)})$ resembles a penalized weighted least squares regression, so algorithms used for maximizing penalized least squares can be applied. Such algorithms include the local quadratic approximation algorithm (LQA) (Fan and Li (2001)), the best convex minorization-maximization algorithm (MM) (Hunter and Li (2005)), and the local linear approximation algorithm (LLA) (Zou and Li (2008)). We use the local linear approximation method to maximize $Q_{1,\lambda}(\boldsymbol{\beta},\boldsymbol{\tau}^{(s)}|\boldsymbol{\eta}^{(s)})$, because it has been shown to reduce the computational cost of maximizing penalized likelihoods (Zou and Li (2008)). Even though an approximation is used for $Q_{1,\lambda}(\boldsymbol{\beta},\boldsymbol{\tau}^{(s)}|\boldsymbol{\eta}^{(s)})$, the maximizer of this function, denoted $\boldsymbol{\beta}^{(s+1)}$, will behave such that $Q_{1,\lambda}(\boldsymbol{\beta}^{(s+1)},\boldsymbol{\tau}^{(s)}|\boldsymbol{\eta}^{(s)}) \geq Q_{1,\lambda}(\boldsymbol{\beta}^{(s)},\boldsymbol{\tau}^{(s)}|\boldsymbol{\eta}^{(s)})$. Therefore, using the ECM algorithm (Meng and Rubin (1993)), we can obtain a $\boldsymbol{\eta}^{(s+1)}$ such that $Q_\lambda(\boldsymbol{\eta}^{(s+1)}|\boldsymbol{\eta}^{(s)}) \geq Q_\lambda(\boldsymbol{\eta}^{(s)}|\boldsymbol{\eta}^{(s)})$, rather than directly maximizing $Q_\lambda(\boldsymbol{\eta}|\boldsymbol{\eta}^{(s)})$. We iterate this process until it converges to a value and denote the value at convergence by $\hat{\boldsymbol{\eta}}_\lambda$. Thus, $\hat{\boldsymbol{\eta}}_\lambda$ maximizes the penalized observed data log-likelihood.

2.3. Penalty selection procedure

To ensure that $\widehat{\boldsymbol{\eta}}_\lambda$ has oracle properties, the penalty parameter $\boldsymbol{\lambda}$ has to be appropriately selected. Two commonly used criteria for selecting the penalty

parameter include the GCV and BIC criteria. These criteria cannot be easily computed in the presence of missing data because they are often functions of the missing data, and thus involve intractable integrals. Moreover, it has been shown that even for the linear model, the GCV can lead to significant overfitting (Wang, Li and Tsai (2007)).

We propose two methods to select the penalty parameter: an IC_Q criterion and a random effects penalty estimation method. The IC_Q criterion selects the optimal $\boldsymbol{\lambda}$ by minimizing

$$\mathrm{IC}_Q(\boldsymbol{\lambda}) = -2Q(\widehat{\boldsymbol{\eta}}_\lambda|\widehat{\boldsymbol{\eta}}_0) + \hat{c}_n(\widehat{\boldsymbol{\eta}}_\lambda),$$

where $\widehat{\boldsymbol{\eta}}_0 = \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \sum_{i=1}^n \log f(\mathbf{D}_{o,i}|\boldsymbol{\eta})$ is the unpenalized maximum likelihood estimate under the full model, and $\hat{c}_n(\boldsymbol{\eta})$ is a function of the data and the fitted model. For instance, if $\hat{c}_n$ equals twice the total number of parameters, then we obtain an AIC-type criterion; alternatively, we obtain a BIC-type criterion when $\hat{c}_n(\boldsymbol{\eta}) = \dim(\boldsymbol{\eta}) \times \log n$. Moreover, in the absence of missing data, we just obtain the usual AIC or BIC criteria. In practice, it is easy to compute IC_Q for different $\boldsymbol{\lambda}$ because we only need samples from $f(\mathbf{z}_{m,i}|y_i, \mathbf{x}_i, \mathbf{z}_{o,i}, \widehat{\boldsymbol{\eta}}_0)$ to approximate $Q(\widehat{\boldsymbol{\eta}}_\lambda|\widehat{\boldsymbol{\eta}}_0)$ at each $\boldsymbol{\lambda}$.

The random effects penalty estimator is calculated under the assumption that the regression coefficients $\boldsymbol{\beta}$ are distributed as random effects in a hierarchical model. The parameter $\boldsymbol{\lambda}$ can be regarded as a parameter in the distribution of $\boldsymbol{\beta}$, denoted by $f(\boldsymbol{\beta}|\boldsymbol{\lambda}, n)$. Then, $\boldsymbol{\lambda}$ can be estimated by maximizing the marginal likelihood given by

$$\int \prod_{i=1}^n \int f(y_i, \mathbf{z}_i, \mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\eta}) f(\boldsymbol{\beta}|\boldsymbol{\lambda}, n) d\mathbf{z}_{m,i} d\boldsymbol{\beta} = \prod_{i=1}^n \int f(\mathbf{D}_{o,i}|\boldsymbol{\eta}) f(\boldsymbol{\beta}|\boldsymbol{\lambda}, n) d\boldsymbol{\beta}, \quad (2.4)$$

where

$$f(\boldsymbol{\beta}|\boldsymbol{\lambda}, n) = \prod_{j=1}^p \exp \frac{-np_{\lambda_j}(|\beta_j|)}{[C(\lambda_j, n)]^p}, \quad (2.5)$$

in which $C(\lambda_j, n)$ is the normalizing constant of $\exp(-np_{\lambda_j}(|\beta_j|))$. The resulting estimate of $\boldsymbol{\lambda}$, denoted by $\hat{\boldsymbol{\lambda}}_{RE}$, from the maximization of (2.4) is the random effects penalty estimator. The EM algorithm can be used to calculate $\hat{\boldsymbol{\lambda}}_{RE}$ by treating the regression coefficients as missing data in the marginal likelihood.

We consider the SCAD and ALASSO penalties as follows. For ALASSO,

$$p_{\lambda_j}(|\beta_j|) = \lambda_j|\beta_j|$$

for $j = 1, \ldots, p$. Typical values chosen are $\lambda_j = \lambda_0|\hat{\beta}_j|^{-\gamma}$, where $\hat{\beta}_j$ is the unpenalized ML estimate and $\gamma > 0$ is a pre-specified positive scalar. In contrast,

the SCAD penalty (Fan and Li (2001)) is a nonconcave function defined by $p_\lambda(0) = 0$ and for $|\beta| > 0$,

$$p'_\lambda(|\beta|) = \lambda 1(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{a-1} 1(|\beta| > \lambda),$$

where $1(\cdot)$ denotes the indicator function, t_+ denotes the positive part of t, and $a = 3.7$. Because the function $\exp(-np_\lambda(|\beta|))$ for the SCAD penalty is not proper, we use a truncated version of $p_\lambda(|\beta|)$ to define the density $f(\boldsymbol{\beta}|\boldsymbol{\lambda}, n)$. For SCAD, we have

$$f(\beta|\lambda, n)C(\lambda, n) = \begin{cases} \exp(-n\lambda|\beta|), & |\beta| < \lambda, \\ \exp\left(\dfrac{n[|\beta|^2 - 2a\lambda|\beta| + \lambda^2]}{[2(a-1)]}\right), & \lambda \le |\beta| \le a\lambda, \\ \exp\left(\dfrac{-n(a+1)\lambda^2}{2}\right), & a\lambda \le |\beta| \le |\bar{\beta}|, \\ 0, & |\beta| > |\bar{\beta}|, \end{cases}$$

where $\bar{\boldsymbol{\beta}}$ is arbitrarily large. For the ALASSO penalty, this truncation is not necessary because $\exp(-np_\lambda(|\beta|))$ is proper.

A closed form expression of $\hat{\boldsymbol{\lambda}}_{RE}$ is unavailable for both the ALASSO and SCAD penalties. But for the ALASSO penalty, a closed form expression of the conditional maximizer of the log-likelihood function with respect to $\boldsymbol{\lambda}$ is available. This allows a straightforward implementation of the ECM algorithm to estimate $\boldsymbol{\lambda}$. For the SCAD penalty, we use the Newton Raphson algorithm along with the ECM algorithm to estimate $\hat{\boldsymbol{\lambda}}_{RE}$.

3. Theoretical Results

In this section, we establish the asymptotic theory of penalized likelihood estimators and the consistency of the penalty selection procedure based on IC_Q. Suppose that $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T)^T$, where $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$ are, respectively, $p_1 \times 1$ and $p_2 \times 1$ subvectors. Let $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_{(1)}^{*T}, \boldsymbol{\beta}_{(2)}^{*T})^T$ denote the true value of $\boldsymbol{\beta}$. Without loss of generality, we assume that $\boldsymbol{\beta}_{(2)}^* = 0$ and each of the components of $\boldsymbol{\beta}_{(1)}$ is not zero.

Let $\mathcal{S} = \{j_1, \ldots, j_d\}$ be a candidate model containing the j_1th, …, j_dth covariates. Thus, $\mathcal{S}_F = \{1, \ldots, p\}$ and $\mathcal{S}_T = \{1, \ldots, p_1\}$ denote the full and true covariate models, respectively. If $\mathcal{S}$ misses at least one important covariate, $\mathcal{S} \not\supset \mathcal{S}_T$, then $\mathcal{S}$ is referred to as an underfitted model; however, if $\mathcal{S} \not\supset \mathcal{S}_T$, then $\mathcal{S}$ is an overfitted model. Assume that we only consider the selected covariates

in $\mathcal{S}$. The unpenalized and penalized ML estimates of $\boldsymbol{\eta}$, denoted by $\widehat{\boldsymbol{\eta}}_S$ and $\widehat{\boldsymbol{\eta}}_\lambda$, respectively, are

$$\widehat{\boldsymbol{\eta}}_S = \underset{\boldsymbol{\eta}:\beta_j\neq 0,\ \forall j\in\mathcal{S}}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(\mathbf{D}_{0,i}|\boldsymbol{\eta}) \ \text{ and } \ \widehat{\boldsymbol{\eta}}_\lambda = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\ P(\boldsymbol{\eta}|\lambda),$$

where $\widehat{\boldsymbol{\eta}}_{\mathcal{S}_F} = \widehat{\boldsymbol{\eta}}_0$.

Theorem 1. *Under assumptions* (C1)−(C7) *stated in the online supplement, we have*

(i) $\widehat{\boldsymbol{\eta}}_\lambda - \boldsymbol{\eta}^* = O_p(n^{-1/2})$ *as* $n\to\infty$, *where* $\widehat{\boldsymbol{\eta}}_\lambda = \left(\hat{\boldsymbol{\beta}}_{(1)\lambda}^T, \hat{\boldsymbol{\beta}}_{(2)\lambda}^T, \hat{\boldsymbol{\tau}}_\lambda^T, \hat{\boldsymbol{\alpha}}_\lambda^T, \hat{\boldsymbol{\xi}}_\lambda^T\right)^T$ *and* $\boldsymbol{\eta}^*$ *is the true value of* $\boldsymbol{\eta}$.

(ii) *Sparsity:* $P(\hat{\boldsymbol{\beta}}_{(2)\lambda} = 0) \to 1$.

(iii) *Asymptotic normality:* $(\hat{\boldsymbol{\beta}}_{(1)\lambda}^T, \hat{\boldsymbol{\tau}}_\lambda^T, \hat{\boldsymbol{\alpha}}_\lambda^T, \hat{\boldsymbol{\xi}}_\lambda^T)^T$ *is asymptotically normal with mean and covariance defined in the online supplement.*

The proof of Theorem 1 is given in the online supplement at `http://www.stat.sinica.edu.tw/statistica`. It states that, by choosing the penalty $\boldsymbol{\lambda}$, there exists a root-n estimator of $\boldsymbol{\eta}$, $\widehat{\boldsymbol{\eta}}_\lambda$, and that this estimator must posses the sparsity property, i.e., $\hat{\boldsymbol{\beta}}_{(2)\lambda} = 0$. Theorem 1(iii) has $\widehat{\boldsymbol{\eta}}_\lambda$ asymptotically normal. An expression for the asymptotic covariance matrix of $\widehat{\boldsymbol{\eta}}_\lambda$ can be obtained using Louis's method (Louis (1983)). These estimates are given in the online supplement.

We investigate whether the $\mathrm{IC}_Q(\lambda)$ criterion can consistently select the correct model. For each $\boldsymbol{\lambda}\in R^{p+}$, $\hat{\boldsymbol{\beta}}_\lambda$ naturally defines a candidate model $\mathcal{S}_\lambda = \{j : \hat{\boldsymbol{\beta}}_{\lambda,j}\neq 0\}$. Generally, $\mathcal{S}_\lambda$ can be either underfitted, overfitted, or true. Therefore, R^{p+} can be partitioned into three mutually exclusive regions $R_u^{p+} = \{\boldsymbol{\lambda}\in R^{p+} : \mathcal{S}_\lambda \not\supset \mathcal{S}_T\}$, $R_t^{p+} = \{\boldsymbol{\lambda}\in R^{p+} : \mathcal{S}_\lambda = \mathcal{S}_T\}$, and $R_o^{p+} = \{\boldsymbol{\lambda}\in R^{p+} : \mathcal{S}_\lambda \supset \mathcal{S}_T, \mathcal{S}_\lambda \neq \mathcal{S}_T\}$. Furthermore, we can always choose a reference penalty parameter sequence $\{\boldsymbol{\lambda}_n \in R^{p+}\}_{n=1}^\infty$, that satisfies the conditions necessary for Theorem 1 to hold. Thus, $\mathcal{S}_{\lambda_n} = \mathcal{S}_T$ with probability converging to one. To select a better model, we first calculate

$$\mathrm{dIC}_Q(\boldsymbol{\lambda}_2,\boldsymbol{\lambda}_1) = \mathrm{IC}_Q(\boldsymbol{\lambda}_2) - \mathrm{IC}_Q(\boldsymbol{\lambda}_1) = 2Q(\widehat{\boldsymbol{\eta}}_{\lambda_1}|\widehat{\boldsymbol{\eta}}_0) - \hat{c}_n(\widehat{\boldsymbol{\eta}}_{\lambda_1}) - 2Q(\widehat{\boldsymbol{\eta}}_{\lambda_2}|\widehat{\boldsymbol{\eta}}_0) + \hat{c}_n(\widehat{\boldsymbol{\eta}}_{\lambda_2}).$$

We assume $\mathcal{S}_{\lambda_2} \supset \mathcal{S}_{\lambda_1}$ and choose the model resulting from using the penalty value $\boldsymbol{\lambda}_1$ (i.e., $\mathcal{S}_{\lambda_1}$), if $\mathrm{dIC}_Q(\boldsymbol{\lambda}_2,\boldsymbol{\lambda}_1)\geq 0$, otherwise we choose model $\mathcal{S}_{\lambda_2}$.

Define $\delta_Q(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2) = E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_{\lambda_1}}|\boldsymbol{\eta}^*)] - E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_{\lambda_2}}|\boldsymbol{\eta}^*)]$, and $\delta_c(\boldsymbol{\lambda}_2,\boldsymbol{\lambda}_1) = \hat{c}_n(\widehat{\boldsymbol{\eta}}_{\lambda_2}) - \hat{c}_n(\widehat{\boldsymbol{\eta}}_{\lambda_1})$, in which $\boldsymbol{\eta}^*_{\mathcal{S}}$ is defined in the online supplement.

Theorem 2. *Under assumptions* (C1)−(C7) *in the Appendix of the online supplement, we have following results.*

(a) *If for all* $\mathcal{S}_\lambda \not\supset \mathcal{S}_T$, $\liminf_n \delta_Q(\boldsymbol{\lambda}, 0)/n > 0$ *and* $\delta_c(\boldsymbol{\lambda}, 0) = o_p(n)$, *then* $dIC_Q(\boldsymbol{\lambda}, 0) > 0$ *in probability for all* $\mathcal{S}_\lambda \not\supset \mathcal{S}_T$.

(b) *If* $E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_{\lambda_1}}|\widehat{\boldsymbol{\eta}}_0)] - E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_{\lambda_2}}|\widehat{\boldsymbol{\eta}}_0)] = O_p(n^{1/2})$ *and* $Q(\widehat{\boldsymbol{\eta}}_{\lambda_t}|\widehat{\boldsymbol{\eta}}_0) - E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_{\lambda_t}}|\widehat{\boldsymbol{\eta}}_0)] = O_p(n^{1/2})$ *for* $t = 1, 2$, *then* $dIC_Q(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) > 0$ *in probability as* $n^{-1/2}\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) \xrightarrow{p} \infty$.

(c) *If* $Q(\widehat{\boldsymbol{\eta}}_{\lambda_1}|\widehat{\boldsymbol{\eta}}_0) - Q(\widehat{\boldsymbol{\eta}}_{\lambda_2}|\widehat{\boldsymbol{\eta}}_0) = O_p(1)$, *then* $dIC_Q(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) > 0$ *in probability as* $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) \xrightarrow{p} \infty$.

The proof of Theorem 2 is given in the online supplement. Theorem 2 has some important implications. Theorem 2a shows that $\mathrm{IC}_Q(\boldsymbol{\lambda})$ chooses all significant covariates with probability 1. Because $\mathcal{S}_0 \subset R_t^p \cup R_o^p$, the optimal model selected when minimizing $\mathrm{IC}_Q(\boldsymbol{\lambda})$ will not select a $\boldsymbol{\lambda}$ with $\mathcal{S}_\lambda \not\supset \mathcal{S}_T$ because $\mathrm{dIC}_Q(\boldsymbol{\lambda}, 0) > 0$ in probability. Therefore, IC_Q selects all significant covariates with probability tending to 1. Generally, the most commonly used $\hat{c}_n(\boldsymbol{\eta})$, such as $2\dim(\boldsymbol{\eta})$, $\dim(\boldsymbol{\eta})\log(n)$, and $K\log\log(n)$ $(K > 0)$, satisfy the condition $\delta_c(\boldsymbol{\lambda}, 0) = o_p(n)$. The condition $\liminf_n n^{-1}\delta_Q(\boldsymbol{\lambda}, 0) > 0$ ensures that $\mathrm{IC}_Q(\boldsymbol{\lambda})$ chooses a model with large $E[Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)]$. This condition is analogous to Condition 2 in Wang, Li and Tsai (2007), which elucidates the effect of models that underfit. Because $n^{-1}E[Q(\boldsymbol{\eta}^*|\boldsymbol{\eta}^*)] - n^{-1}E[Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)]$ can be written as

$$n^{-1}\sum_{i=1}^n \log f(\mathbf{D}_{o,i}|\boldsymbol{\eta}^*) - n^{-1}\sum_{i=1}^n \log f(\mathbf{D}_{o,i}|\boldsymbol{\eta}^*_{\mathcal{S}}) + n^{-1}E\Big[H(\boldsymbol{\eta}^*|\boldsymbol{\eta}^*)\Big] - n^{-1}E\Big[H(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)\Big],$$

where

$$H(\boldsymbol{\eta}|\boldsymbol{\eta}_1) = \int \sum_{i=1}^n \log\Big[f(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i, \mathbf{r}_i, \boldsymbol{\eta})\Big] f(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i, \mathbf{r}_i, \boldsymbol{\eta}_1) d\mathbf{z}_{m,i},$$

it then follows from Jensen's inequality that $n^{-1}\delta_Q(\lambda, 0) \geq 0$. Thus, if a model $\mathcal{S}$ misses a significant covariate, it is reasonable to assume $\liminf_n n^{-1}\delta_Q(\boldsymbol{\lambda}, 0)$ is greater than zero.

If $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ have the same average $n^{-1}E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_\lambda}|\boldsymbol{\eta}^*)]$, that is, $\liminf_n n^{-1}\delta_Q(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) = 0$, then Theorem 2 (b) and (c) indicate that $\mathrm{IC}_Q(\boldsymbol{\lambda})$ picks out the smaller model $\mathcal{S}_{\lambda_1}$ when $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1)$ increases to ∞ at a certain rate (e.g., $\log(n)$). For example, for the BIC-type criterion, $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) = [\dim(\widehat{\boldsymbol{\eta}}_{\mathcal{S}_{\lambda_2}}) - \dim(\widehat{\boldsymbol{\eta}}_{\mathcal{S}_{\lambda_1}})]\log(n) \geq \log(n)$, since we assume $\mathcal{S}_{\lambda_2} \supset \mathcal{S}_{\lambda_1}$. However, the AIC-type criterion $\hat{c}_n(\boldsymbol{\eta}) = 2 \times \dim(\boldsymbol{\eta})$ does not satisfy this condition. Thus, similar to the standard AIC, IC_Q with $\hat{c}_n(\boldsymbol{\eta}) = 2 \times \dim(\boldsymbol{\eta})$ tends to overfit.

4. Numerical Studies

4.1. Example 1: simulation study

We demonstrate the performance of the penalized ML estimates using our proposed penalty estimators via simulations and compare them to the unpenalized ML estimate. Our objective for these simulations was to (1) compare the performance of the random effects and the IC_Q penalty estimators, (2) compare the performance of the SCAD and ALASSO penalty functions, and (3) determine how the comparisons in (1) and (2) differ in the complete data and missing covariate settings.

To do this, we simulated datasets consisting of n observations from the model $y = \mathbf{u}^T\boldsymbol{\beta}^* + \sigma\epsilon$ where $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and the components of $\mathbf{u} = (u_1, \ldots, u_8)$, and ϵ are standard normal. The correlation between u_i and u_j is $\rho^{|i-j|}$ with $\rho = 0.5$. This model was used in Fan and Li (2001). We considered three settings, $(n = 40, \sigma = 3)$, $(n = 40, \sigma = 1)$, and $(n = 60, \sigma = 1)$. For each of them, two sets of 100 datasets were simulated, one with complete data and another with missing covariate data. For the datasets with missing data, the missing covariates $\mathbf{z}_i = (u_{1i}, u_{2i})$ were taken to be MAR and $\mathbf{x}_i = (u_{3i}, \ldots, u_{8i})$ were completely observed. The covariate distribution is given by, $[\mathbf{z}_i|\mathbf{x}_i] \sim N_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$ where $\boldsymbol{\mu}_i = (\mu_{1i}, \mu_{2i})$, $\mu_{si} = \alpha_{s0} + \sum_{j=1}^{5} \alpha_{sj} x_{is}$ for $s = 1, 2$ and $\boldsymbol{\Sigma}$ is an unstructured 2×2 covariance matrix. The missing data mechanism used was $f(r_{i1}, r_{i2}|y_i, \mathbf{x}_i, \boldsymbol{\phi}) = f(r_{i1}|r_{i2}, y_i, \mathbf{x}_i, \boldsymbol{\phi}_1) f(r_{i2}|y_i, \mathbf{x}_i, \boldsymbol{\phi}_2)$, where $f(r_{i1}|y_i, x_i, \boldsymbol{\phi}_1)$ and $f(r_{i2}|r_{i1}, y_i, x_i, \boldsymbol{\phi}_2)$ are logistic regressions where the logistic regression parameters $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ were selected such that 65% of the observations had complete data.

For each simulated dataset, the penalized ML estimate using the SCAD and ALASSO penalties was computed using the random effects and IC_Q penalty estimates. These estimates are denoted as SCAD-RE, SCAD-IC_Q, ALASSO-RE, and ALASSO-IC_Q, respectively. For the IC_Q estimate, the BIC-type criterion, $c_n(\boldsymbol{\eta}) = \dim(\boldsymbol{\eta}) \log n$, was used. In the analysis of the datasets with no missing covariates, the IC_Q criterion is equivalent to BIC. For the random effects penalty estimator, 2,000 Monte Carlo iterations were used within each iteration of EM. Since the EM algorithm can be sensitive to starting values, the algorithm was initiated from multiple starting values to ensure the overall global maximum was achieved by the algorithm. For the ALASSO penalty, we set $\lambda_j = \lambda_0 |\widehat{\beta}_{j0}|^{-1}$, where $\hat{\beta}_{j0}$ is the unpenalized ML estimate and for the SCAD penalty we let $\lambda_j = \lambda_0$, for all j, where in both cases λ_0 was estimated using the penalty estimation methods.

In addition to the penalized estimates, the unpenalized ML estimate of the model selected by the simultaneously impute and select (SIAS) method of Yang, Belin and Boscardin (2005) was computed. SIAS implements the stochastic search variable selection (SSVS) method of George and McCulloch (1993) in the presence of missing covariates. SIAS is a fully Bayesian method which does not require model enumeration or computation of marginal likelihoods, so it maybe easier to implement than other fully Bayesian methods. In the analysis of the datasets with no missing covariates, SIAS is equivalent to SSVS. Details of the implementation of SIAS are given in the online supplement.

For each estimate $\hat{\boldsymbol{\beta}}_\lambda$, the model error, $\mathrm{ME}(\hat{\boldsymbol{\beta}}_\lambda) = (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)E(\mathbf{u}\mathbf{u}^T)(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)$, was computed and the ratio of the model error of the penalized ML estimate to that of the unpenalized ML estimate, $\mathrm{ME}(\hat{\boldsymbol{\beta}}_\lambda)/\mathrm{ME}(\hat{\boldsymbol{\beta}}_0)$, was computed. The median of these ratios over the 100 simulated datasets, denoted as MRME, is reported. The MRME of the true model, denoted as 'oracle', is also reported. In addition, the average number of zero coefficients correctly estimated to be zero and the average number of zero coefficients incorrectly estimated to be zero are reported. These are reported in the columns 'Correct' and 'Incorrect' respectively.

The results indicate that when the noise level is high ($\sigma = 3$), the ALASSO-RE and SCAD-IC$_Q$ estimates have smallest model error while the SCAD-RE has the highest. When the noise level is reduced ($\sigma = 1$), or the sample size is large ($n = 60$), the SCAD-RE estimate has the smallest model error. For the estimates, MRME values greater than one indicate that the estimate performs worse than the unpenalized ML estimate, values near one indicate it performs as good as the unpenalized ML estimate, while values near the 'oracle' MRME value indicate optimal performance. The SCAD-RE performed poorly when the noise level was high, however, it is optimal when either the noise level is small or the sample size is large. The ALASSO-RE estimate had substantial overfit since 'Correct' averaged significantly less than 5 indicating a tendency to not set insignificant coefficients to zero. The SIAS estimate performed as well as the unpenalized ML estimate when the noise level was large and covariates were missing, however it outperformed the ML estimate when either the noise level was high, the sample size was large, or all the covariates were fully observed. 'Correct' averages and 'Incorrect' averages that are both high indicate that the estimate is more likely to set coefficients to zero rather than not. This was the case with the SIAS and SCAD-RE estimates when the noise level was large. Comparing the analysis of no missing covariate data to the analysis with missing covariate data shows that for all the estimates, the estimation error increased, overfitting increased, and underfitting increased.

Table 4.1. Simulation results of linear regression model with no missing data and covariates missing at random comparing SCAD and ALASSO penalty functions with random effects and IC_Q penalty estimates.

	No missing (MAR)			
			# of 0 coefficients	
Model	Method	MRME	Correct	Incorrect
$n = 40$, $\sigma = 3$	SCAD-RE	1.111 (1.203)	4.91 (4.90)	0.97 (0.98)
	SCAD-IC_Q	0.625 (0.745)	4.53 (4.48)	0.33 (0.45)
	ALASSO-RE	0.632 (0.690)	3.23 (3.42)	0.09 (0.13)
	ALASSO-IC_Q	0.681 (0.771)	4.31 (4.23)	0.28 (0.35)
	SIAS	0.765 (1.004)	4.81 (4.87)	0.55 (0.77)
	Oracle	0.256 (0.305)	5.00 (5.00)	0.00 (0.00)
$n = 40$, $\sigma = 1$	SCAD-RE	0.285 (0.316)	4.34 (4.49)	0.01 (0.01)
	SCAD-IC_Q	0.333 (0.549)	4.64 (4.15)	0.00 (0.00)
	ALASSO-RE	0.472 (0.543)	3.45 (3.23)	0.00 (0.00)
	ALASSO-IC_Q	0.404 (0.572)	4.58 (4.10)	0.00 (0.00)
	SIAS	0.321 (0.360)	4.82 (4.79)	0.00 (0.00)
	Oracle	0.273 (0.258)	5.00 (5.00)	0.00 (0.00)
$n = 60$, $\sigma = 1$	SCAD-RE	0.322 (0.351)	4.54 (4.62)	0.00 (0.00)
	SCAD-IC_Q	0.375 (0.386)	4.86 (4.73)	0.00 (0.00)
	ALASSO-RE	0.517 (0.495)	3.47 (3.53)	0.00 (0.00)
	ALASSO-IC_Q	0.425 (0.447)	4.83 (4.70)	0.00 (0.00)
	SIAS	0.461 (0.387)	4.70 (4.82)	0.00 (0.00)
	Oracle	0.310 (0.356)	5.00 (5.00)	0.00 (0.00)

4.3. Example 2: melanoma data

To further illustrate our proposed methods, we consider data on $n = 286$ patients from a phase III two arm clinical trial conducted by the Eastern Cooperative Oncology Group. The results from this study have been reported in Kirkwood, Strawderman, Ernstoff, Smith, Borden and Blum (1996). Patients in this trial were randomized to one of two treatment arms: high dose interferon or observation. Interferon is suggested to have a significant effect on disease-free survival. Here, disease free survival is defined as the time from randomization until progression of tumor or death, whichever comes first. In this analysis, several prognostic factors were identified as important predictors of survival. Among these factors are, z_1 = Breslow thickness (in mm), z_2 = size of primary (in cm^2), z_3 = type of primary tumor (two levels: superficial spreading, other), x_1 = age (in years), x_2 = pathological group (two levels: previous recurrence and other) and x_3 = treatment (two levels: high dose interferon and observation). From these six covariates, three had missing data while the rest of the covariates and the response variable were completely observed. The three covariates with missing data were Breslow thickness, size, and type. Logarithms of Breslow thickness

and size were used in this analysis to achieve approximate normality of these covariates in the covariate distribution. The dataset had a total missing data fraction of 28.7%. The outcome variable, y_i, was taken here to be binary, and was assigned a 1 if the patient had an overall survival greater than or equal to 0.55 years, and 0 otherwise. There were no censored cases that had an overall survival below 0.55 years.

To analyze these data, a logistic regression model was used for $y_i|\mathbf{x}_i, \boldsymbol{\beta}$ with $E(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \exp(\gamma_i)/(1+\exp(\gamma_i))$, where $\gamma_i = (1, \mathbf{z}_i, \mathbf{x}_i)^T\boldsymbol{\beta}$, $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})^T$, $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^T$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_6)$. For the missing covariates, we assume they are MAR and have the covariate distribution

$$f(\mathbf{z}_i|\mathbf{x}_i; \boldsymbol{\alpha}) = f(z_{i3}|z_{i1}, z_{i2}, \mathbf{x}_i; \boldsymbol{\alpha}_3) f(z_{i1}, z_{i2}|\mathbf{x}_i; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$$

for $i = 1, \ldots, n$. Since $\mathbf{x}_i$ is completely observed, it is conditioned on throughout. We take $(z_{i1}, z_{i2}|\mathbf{x}_i) \sim \boldsymbol{N}_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2})$ and $\mu_{is} = \alpha_{s0} + \sum_{j=1}^{3} \alpha_{sj} x_{ij}$ for $s = 1, 2$, $i = 1, \ldots, n$, and $\boldsymbol{\Sigma}$ is an unstructured 2×2 covariance matrix. A logistic regression model was used for x_{i3} conditional on $(z_{i1}, z_{i2}, \mathbf{x}_i)$. The same estimates as those computed in the simulations were computed. The statistical model used for the SIAS method is given in the online supplement.

The results are presented in Table 4.2. The predictors identified as significant were different for the each of the estimation methods. In the missing data analysis, the ALASSO and SIAS estimates identified treatment as a significant predictor while the SCAD estimates did not. The ALASSO-IC$_Q$ estimate also identified treatment and pathology as significant while the ALASSO-RE estimate identified treatment, pathology and age as significant. According to the unpenalized ML analysis, treatment and pathology are the only predictors which are possibly significant since their p-values are near or below the cutoff value of 0.05 for significance. However, neither of these predictors was strongly significant. Therefore, a possible explanation for the differences in the results of the various estimation methods is that these methods may not be able to discriminate between models that include or exclude treatment and pathology very well. The results of the unpenalized maximum likelihood analysis coincided with the results of the ALASSO-IC$_Q$ and SIAS estimates. As with the simulations, the ALASSO-RE estimate tended to overfit since it identified age as significant even though its p-value was greater than 0.05, and the SCAD-RE estimate tended to set coefficients to 0 since it did not identify any predictors as significant. The estimate of the regression coefficient for treatment decreased from 1.117 in the complete case analysis to 0.839 in the missing data analysis. This change caused the SCAD-IC$_Q$ estimate to identify treatment as significant in the complete case analysis but not significant for the missing data analysis.

Table 4.2. Estimates of Melanoma Data

	Missing Data Estimate						
	SCAD		ALASSO		SIAS	MLE (p value)	
Variable	RE	IC_Q	RE	IC_Q			
Intercept	2.132	2.132	2.421	2.280	1.774	2.638	(<0.001)
Breslow	0.000	0.000	0.000	0.000	0.000	-0.217	(0.332)
Size	0.000	0.000	0.000	0.000	0.000	-0.052	(0.798)
Type	0.000	0.000	0.000	0.000	0.000	-0.161	(0.730)
Age	0.000	0.000	-0.267	0.000	0.000	-0.325	(0.146)
Pathology	0.000	0.000	-0.845	-0.454	0.000	-1.061	(0.039)
Treatment	0.000	0.000	0.737	0.322	0.827	0.839	(0.043)
	Complete Case Estimate						
	SCAD		ALASSO		SIAS	MLE (p value)	
Variable	RE	IC_Q	RE	IC_Q			
Intercept	2.085	1.609	2.043	1.820	1.609	2.210	(<0.001)
Breslow	0.000	0.000	-0.081	0.000	0.000	-0.222	(0.400)
Size	0.000	0.000	0.000	0.000	0.000	-0.089	(0.650)
Type	0.000	0.000	0.000	0.000	0.000	0.235	(0.650)
Age	0.000	0.000	-0.113	0.000	0.000	-0.232	(0.356)
Pathology	0.000	0.000	-0.578	0.000	0.000	-0.945	(0.086)
Treatment	0.000	1.173	1.003	0.572	1.173	1.117	(0.028)

5. Discussion

We have proposed a general method to simultaneously perform model selection and estimation in the presence of missing data. We have showed that under regularity conditions and appropriate rates of the penalty parameter, the penalized estimate possesses oracle properties. We have introduced two computationally attractive methods for estimating the penalty parameters. We have showed that under an appropriate choice of $\hat{c}_n(\boldsymbol{\eta})$, the IC_Q penalty estimate chooses all the significant predictors in probability. Simulation results show that the SCAD penalty function with the random effects penalty estimate performs well when the noise level is small, whereas it performs poorly when the noise level is large. Overall, the SCAD performed better when it was used with the random effects penalty estimator whereas the ALASSO performed better when it was used with the IC_Q criterion. The ALASSO penalty function with the random effects penalty estimate showed significant overfit in the finite sample simulations and this overfit was also present in the Melanoma data analyses. The results of the Melanoma data analysis indicate that when predictors are not strongly significant, the results from penalized likelihood maximization may differ depending on the penalty functions and penalty selection methods which are used.

One of the disadvantages of penalized likelihood methods is that they do not provide a measure of model uncertainty, i.e., the probability of selecting each

model in the model space. Other methods, such as Bayesian model averaging (Hoeting, Madigan, Raftery and Volinsky (1999)), SIAS, or Bayesian methods in general provide estimates of posterior model probabilities. However, implementation of fully Bayesian methods can be difficult in many cases, since it requires specifying priors for all of the parameters in the response model, covariate distribution (and missing data mechanism under NMAR) which encompass all the models in the model space, as well as calculating marginal likelihoods and enumerating all the models in the model space. Alternatively, the SIAS method is easier to implement but, unlike penalized ML maximization, it does not give an estimate of the parameters of the 'best' model. Moreover, the results of the linear regression simulations indicated that the SCAD-RE estimate outperforms SIAS when either the noise level is small or the sample size is large.

Many aspects of this work warrant further research and investigation. One major issue is to carry out variable selection using IC_Q under different modeling situations such as generalized linear mixed models with nonignorable missing response and/or covariate data, semiparametric survival models with missing covariate data, such as the Cox model as well as frailty models, measurement error models, and partially linear models with missing covariates and/or responses. Throughout this paper, we made an implicit assumption that the response model does not depend on whether a covariate is observed or missing. That is, we have assumed a single response model for the covariate where it is missing or not. If we have a different response model for the observed and missing parts of the covariate, then the methods developed in this paper would not be able detect whether the missing part of a covariate is significant. In this scenario other statistical methods, such as propensity score methods, may be useful for handling this case (Kang and Schafer (2007)), but applying these methods to variable selection problems requires further developments both computationally and theoretically. We will formally investigate these issues in our future work.

References

Bickel, P. J. and Li, B. (2006). Regularization in statistics. *Test* **76**, 271-344.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99**, 710-723.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881-889.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial, *Statist. Sci.* **14**, 382-417.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617-1642.

Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55**, 591-596.

Ibrahim, J. G. and Lipsitz, S. R. (1996). Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics* **52**, 1071-1078.

Ibrahim, J. G., Lipsitz, S. R. and Chen, M. H. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika* **88**, 551-564.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparision of alternative strategies from estimating a population mean from incomplete data. *Statist. Sci.* **22**, 523-539.

Kirkwood, J. M., Strawderman, M. H., Ernstoff, M. S., Smith, T. J., Borden, E. C. and Blum, R. H. (1996). Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the eastern cooperative oncology group trial EST 1684. *Journal of Clinical Oncology* **14**, 7-17.

Little, R. J. A. and Schluchter, M. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 497-512.

Louis, T. A. (1983). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226-233.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall, London.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267-78.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selector for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568,

Yang, X., Belin, T. R. and Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics* **61**, 498-506.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Statist. Sci.* **36**, 1509-1533.

Department of Biostatistics, University of North Carolina School of Public Health, McGavran Greenberg Hall, Chapel Hill, North Carolina 27599, U.S.A.

E-mail: rgarcia@bios.unc.edu

Department of Biostatistics, University of North Carolina School of Public Health, McGavran Greenberg Hall, Chapel Hill, North Carolina 27599, U.S.A.

E-mail: ibrahim@bios.unc.edu

Department of Biostatistics, University of North Carolina School of Public Health, McGavran Greenberg Hall, Chapel Hill, North Carolina 27599, U.S.A.

E-mail: hzhu@bios.unc.edu

(Received March 2008; accepted November 2008)