Variable- and Model-Selection for Multivariate Failure Time Data

Jianwen Cai, Jianqing Fan, Runze Li and Haibo Zhou

Abstract

We discuss a penalized pseudo-partial likelihood method for variable- and model-selection with multivariate failure time data. The proposed method has some nice asymptotic properties. We show that, for certain penalty functions with proper choices of regularization parameters, the resulting estimate is root n consistent and possesses an oracle property, namely, the resulting estimate can correctly identify the true model as if the true model (the subset of variables with nonvanishing coefficients) were known in advance. Using a simple approximation of the penalty function, the proposed method can be easily carried out with the Newton-Raphson algorithm. We conduct extensive Monte Carlo simulation studies to assess the finite sample performance of the proposed procedures. We illustrate the proposed method by analyzing a data set from the Framingham Heart Study.

Key Words and Phrases: Cox's model, Marginal hazards model, penalized likelihood, SCAD, variable selection.

Jianwen Cai is Associate Professor, Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599-7420 (Email: *cai@bios.unc.edu*). Jianqing Fan is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC27599-3260 (Email: *jfan@stat.unc.edu*). Runze Li is Assistant Professor, Department of Statistics, The Pennsylvania State University, University Park, PA 16802-2111 (Email: *rli@stat.psu.edu*). Haibo Zhou is Associate Professor, Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599-7420 (Email: *zhou@bios.unc.edu*). This research was partially supported by NIH grant R01 HL69720. Fan's research was also partially supported by NSF grant DMS-0204329, and a RGC grant CUHK4262/01P of HKSAR. Li's research was supported by NSF grant DMS-0102505 and National Institute on Drug Abuse (NIDA) Grant 1-P50-DA10075.

1 Introduction

Deciding which covariates to be included in the final statistical model has always been a tricky task for investigators facing a large number of covariates in a real study. A common practice in epidemiological studies, for example, is based on the consideration of two factors: a set of important confounding variables choosing *a priori* and other variables choosing with the aid of statistical evidences such as a t-test or Chi-square test. While this art of model selection allows the investigator to integrate their knowledge on subject-matter considerations with the objective measure from statistical testing, it does have limitations in some settings. For example, different models may result from the same study due to different investigators' subjective selection of the *a priori* set. Furthermore, for a given *a priori* set, several non-nested models might fit the data equally well and it is subjective to pick one over the other. A valid and unified statistical model selection criterion is desirable in these situations. We propose a penalized pseudo-partial likelihood method for the variable- and model-selection in multivariate failure time analysis.

Our research is motivated by the need to develop a predictive model between multiple failure time outcomes (time to coronary heart disease and time to cerebrovascular accident) and a vector of risk factors for patients in the Framingham Heart Study (Dawber, 1980). The Framingham Heart Study was a large prospective study designed to identify the common factors or characteristics that contribute to cardiovascular disease by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of cardiovascular disease or suffered a heart attack or stroke. A total of 2873 women and 2336 men were followed since 1948 and they were recalled and examined every two years after their entry into the study. Since the primary sampling unit was the family, it is likely that the failure times recorded for subjects within a family are correlated. The well-known Cox model (Cox, 1972) is not valid in this situation because independence assumption between individuals is violated. Extensions of the Cox regression model for the analysis of multivariate failure time data include the frailty model and the marginal model. The frailty model considers the conditional hazard given the unobservable frailty random variable. It is useful when the correlation between failure times are of interest. Hougaard (2000) gives a comprehensive account on this topic. One of the issues with the frailty model is that the conditional hazard and the marginal hazard cannot satisfy proportional hazards simultaneously, except for the case where the frailty follows the positive stable distribution. The

interpretations of the regression coefficients in the frailty model are different from those in the Cox model. Consequently, when the correlation among the observations is not of interest, the marginal proportional hazard models have received much attention in the recent literature because they are semiparametric models and retain the virtue of the Cox model (e.g., Wei, Lin and Weissfeld 1989, Lee, Wei and Amato 1992, Liang, Self and Chang 1993, Lin 1994, Cai and Prentice 1995, 1997, Spiekerman and Lin 1998 and Clegg, Cai and Sen 1999 among others).

Some of the variable selection criteria and procedures in linear regression analysis have been extended to the Cox model and the frailty model. Tibshirani (1997) extended his LASSO variable selector to the Cox model. Faraggi and Simon (1998) proposed a Bayesian variable selection methods for the Cox model following the idea of Lindley (1968). Fan and Li (2002) extended their nonconcave penalized likelihood approach to the Cox model and the frailty model. Cai (1999) extended the generalized likelihood ratio (GLR) method to the multivariate failure time data and showed that the GLR statistics in this case follows a weighted Chi-square distribution. In general, the variable selection procedure using the penalized likelihood concept under the marginal models for multivariate failure time data is under-developed. This paper intends to fill that gap.

The rest of this paper is organized as follows. In Section 2, we propose a class of variable selection procedures for various marginal models for multivariate failure time data via penalized pseudo-partial likelihood approach. We also present the asymptotic properties. We consider the practical computation aspects in Section 3. The proposed procedure is summarized in an easy-to-follow algorithm. We evaluate the proposed procedure through simulation studies in Section 4 and illustrate the proposed approach via an application to the Framingham Heart Study data set in Section 5. Final remarks are given in Section 6. Regularity conditions and proofs of theoretic results are given in the Appendix.

2 Model Selection with a Penalized Pseudo-partial Likelihood

To fix notation, suppose that there are n independent clusters and each cluster has K subjects. For each subject, J types of failure may occur. For the failure time in the case of the jth type of failure on subject k in cluster i, the marginal hazards model is taken either as

$$h_{ijk}(t|\mathbf{x}_{ijk}(t)) = h_{0j}(t)\exp\{\boldsymbol{\beta}^T\mathbf{x}_{ijk}(t)\}, \tag{2.1}$$

or

$$h_{ijk}(t|\mathbf{x}_{ijk}(t)) = h_0(t)\exp\{\boldsymbol{\beta}^T\mathbf{x}_{ijk}(t)\}, \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^T$ is a vector of unknown regression coefficients, d is the dimension of $\boldsymbol{\beta}$, $\mathbf{x}_{ijk}(t)$ is a possibly external time-dependent covariate vector, and $h_{0j}(t)$ and $h_0(t)$ are unspecified baseline hazard functions. Model (2.1) is referred to as the mixed baseline hazards model, while (2.2) as the common baseline hazards model in the literature (Spiekerman and Lin, 1998 and Clegg, Cai and Sen, 1999). Note that different regression coefficients for different failure types and different subjects in a cluster can be accommondated by (2.1) and (2.2) by introducing subject-specific or failure-specific dummy variables in the covariate vector. For example, model (2.1) is reduced to

$$h_{ijk}(t|\mathbf{x}^0_{ijk}(t)) = h_{0j}(t)\exp\{\boldsymbol{\beta}_j^T\mathbf{x}^0_{ijk}(t)\} \tag{2.3}$$

if we take $\mathbf{x}_{ijk}(t)^T = (I(j=1)\mathbf{x}^0_{ijk}(t)^T, \cdots, I(j=J)\mathbf{x}^0_{ijk}(t)^T)$ and $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_J^T)$, where $I(\cdot)$ is an indicator function.

2.1 Penalized Pseudo-partial Likelihood

The marginal model approach does not specify correlation structure for the failure times within a cluster, hence one cannot make inferences based on the full likelihood or partial likelihood. Rather, inferences are based on a pseudo-partial likelihood approach. For ease of presentation, we drop the subscript and let T, C and $\mathbf{x}(t)$ be the survival time, the censoring time and their associated covariates, respectively. Correspondingly, let $Z = \min\{T, C\}$ be the observed time, $\delta = I(T \le C)$ be the censoring indicator, and $Y(t) = I(Z \ge t)$ be the at-risk indicator. We further assume that T and C are conditionally independent given $\mathbf{x}$ and that the censoring mechanism is noninformative. Under a working independence assumption (Wei, Lin and Weissfield 1989), i.e., assuming the independence among failure times in a cluster, we obtain the natural logarithm of a pseudo-partial likelihood function for model (2.1) as following:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\delta_{ijk}\left(\boldsymbol{\beta}^T\mathbf{x}_{ijk}(Z_{ijk}) - \log\left[\sum_{l=1}^{n}\sum_{g=1}^{K}Y_{ljg}(Z_{ijk})\exp\{\boldsymbol{\beta}^T\mathbf{x}_{ljg}(Z_{ijk})\}\right]\right). \tag{2.4}$$

To balance between modeling biases and estimation variance, many traditional variable selection criteria have resorted to the penalized likelihood, e.g. the AIC (Akaike 1973) and BIC (Schwarz 1978). We use a natural logarithm of penalized pseudo-partial likelihood for model (2.1) which is

defined as

$$\mathcal{L}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{2.5}$$

where $p_{\lambda_j}(|\beta_j|)$ is a given non-negative function called a penalty function with λ_j as a regularization or tuning parameter. For specific choice of penalty functions, maximizing $\mathcal{L}(\boldsymbol{\beta})$ will result in some vanishing estimates of coefficients and their associated variables are deleted. Hence, by maximizing $\mathcal{L}(\boldsymbol{\beta})$, we select a model and estimate its parameters simultaneously. Examples of penalty functions will be given in Section 3. The tuning parameters can be chosen subjectively by data analysts or objectively by data themselves. In general, large values of $\lambda_j's$ result in simpler models with fewer selected variables.

The penalty term in (2.5) is more general than that in Fan and Li (2001) who considered $\lambda_j \equiv \lambda$. Allowing covariate-specific tuning parameters enables different regression coefficients to have different penalty functions and thus, the penalized pseudo-partial likelihood may directly incorporate hierarchical prior information for the unknown coefficients. For instance, we may wish to keep the main effects of some important confounding variables in the model by not penalizing their corresponding coefficients.

2.2 Asymptotic Properties and Oracle Property of the Proposed Estimator

The proposed penalized pseudo-partial likelihood estimator, denoted by $\widehat{\boldsymbol{\beta}}$, is the one that maximizes (2.5). We now present the asymptotic properties for $\widehat{\boldsymbol{\beta}}$ and show that it could perform as well as an oracle estimator. By an oracle estimator, we mean that the estimator is constructed with the aid of an oracle who knows the true model (the subset of variables with nonvanishing coefficients). The oracle property refers to the resulting estimates that can correctly identify the true model as if the true model were known in advance. Denote by $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$. Further let $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\beta}_{20}$ denote the nonzero and zero components of $\boldsymbol{\beta}_0$, respectively. Denote by s the dimension of $\boldsymbol{\beta}_{10}$ and let

$$a_n = \max\{|p'_{\lambda_j}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}, \tag{2.6}$$

and

$$b_n = \max\{|p''_{\lambda_j}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}. \tag{2.7}$$

We first show that there exists a penalized pseudo-partial likelihood estimator that converges at rate $O_P(n^{-1/2} + a_n)$, and then establish the oracle property for the resulting estimator. We only state the main theoretic results here and leave the proofs and regularity conditions in the Appendix.

Theorem 2.1 *Under Conditions A-D in the Appendix, if both a_n and b_n tend to 0 as $n \to \infty$, then with probability tending to one, there exists a local maximizer $\widehat{\boldsymbol{\beta}}$ of $\mathcal{L}(\boldsymbol{\beta})$ defined in (2.5) such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$.*

From Theorem 2.1, provided that $a_n = O(n^{-1/2})$, which can be achieved by choosing proper $\lambda_j's$, there exists a root n consistent penalized pseudo-partial likelihood estimator.

Without loss of generality, we write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$, where $\boldsymbol{\beta}_{10}$ consists of the s nonzero components of $\boldsymbol{\beta}_0$. Denote by

$$\Sigma = \text{diag}\{p''_{\lambda_1}(|\beta_{10}|), \cdots, p''_{\lambda_s}(|\beta_{s0}|)\},$$

and

$$\mathbf{b} = (p'_{\lambda_1}(|\beta_{10}|)\text{sgn}(\beta_{10}), \cdots, p'_{\lambda_s}(|\beta_{s0}|)\text{sgn}(\beta_{s0})).$$

Theorem 2.2 *Assume that the penalty function $p_{\lambda_j}(|\beta_j|)$ satisfies that*

$$\liminf_{n\to\infty} \liminf_{\beta_j \to 0+} p'_{\lambda_j}(\beta_j)/\lambda_j > 0 \tag{2.8}$$

for all $j = 1, \cdots, d$. If $\lambda_j \to 0$, $\sqrt{n}\lambda_j \to \infty$ and $a_n = O(n^{-1/2})$, then under the conditions of Theorem 2.1, with probability tending to 1, the root n consistent local maximizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ in Theorem 2.1 must satisfy that $\widehat{\boldsymbol{\beta}}_2 = 0$, and

$$\sqrt{n}\{A_{11} + \Sigma\}\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (A_{11} + \Sigma)^{-1}\mathbf{b}\} \to N(0, D_{11}) \tag{2.9}$$

in distribution, where A_{11} and D_{11} consist of the first s columns and rows of $A(\boldsymbol{\beta}_{10}, \mathbf{0})$ and $D(\boldsymbol{\beta}_{10}, \mathbf{0})$ defined in the Appendix, respectively.

The above theorem provides a foundation for choosing estimators that will have the oracle property, though proper care is needed in selecting the penalty function (Section 3 provides more details on this). For example, with the SCAD penalty (details in Section 3), we have $a_n = 0$, $\mathbf{b} = 0$ and $\Sigma = 0$ for sufficiently large n. Hence, according to Theorem 2.2, we have

$$\sqrt{n}A_{11}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \to N(0, D_{11}).$$

The estimator $\widehat{\boldsymbol{\beta}}_1$ shares the same sampling property as the oracle estimator. Furthermore, $\widehat{\boldsymbol{\beta}}_2 = 0$ is the same as the oracle estimator who knows in advance $\boldsymbol{\beta}_2 = 0$. In other words, the penalized pseudo-partial likelihood estimator possesses the oracle property. In contrast, it can easily be shown from Theorems 2.1 and 2.2 that the procedure based on the L_1 penalty, defined by $p_\lambda(|\beta|) = \lambda|\beta|$, does not possess the oracle property due to the excessive biases inherent to the L_1 penalty.

Standard error formula

The standard errors for estimated parameters can be directly obtained because we estimate parameters and select variables (via estimating coefficients as zero) simultaneously. Newton-Raphson algorithm will be employed to search the solution of penalized pseudo-partial likelihood. Following the conventional technique in the marginal model for multivariate failure time data, the variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ can be estimated by the sandwich formula (Wei, Lin, and Weissfeld, 1989). The general formula is of the following form,

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}}) = \{\mathcal{L}''(\widehat{\boldsymbol{\beta}})\}^{-1}\widehat{\text{cov}}\{\mathcal{L}'(\widehat{\boldsymbol{\beta}})\}\{\mathcal{L}''(\widehat{\boldsymbol{\beta}})\}^{-1}.$$

Since some penalty functions, such as the SCAD, are singular at the origin, and they do not have continuous second order derivatives, slight modification of the sandwich formula is needed and will be given in Section 3.

3 Computing Algorithm and Other Practical Considerations

Selection of a Penalty Function

Various authors have considered the issue of selection of penalty function in the context of linear regression. The well known L_2 penalty, $p_\lambda(|\beta|) = \frac{1}{2}\lambda|\beta|^2$, leads to a ridge regression to cope with collinearity problem. Classic variable selection criteria are special cases of the penalized likelihood (2.5). For instance, consider the L_0 penalty $p_\lambda(|\beta|) = \frac{1}{2}\lambda^2 I\{|\beta| \neq 0\}$, also called the entropy penalty in the literature, where $I(\cdot)$ is an indicator function. Note that $\sum_j I\{|\beta_j| \neq 0\}$ equals to the number of regression coefficient of an underlying model. Thus, with the entropy penalty, many variable selection criteria can be written in the form of (2.5). For example, AIC (Akaike 1973) and BIC (Schwarz 1978) correspond to $\lambda = \sqrt{2/n}$ and $\sqrt{\log(n)/n}$, respectively, where n is the sample size. Shao (1997) studied asymptotic consistency of a class of penalized least squares with the

entropy penalty, including GIC (Rao and Wu 1989) and its analogues (Pötscher (1989) and Shao and Rao (2000)). Since the entropy penalty function is discontinuous, it requires searching over all possible subsets for finding the solution. Namely, find the best subset of J variables and then choose J to optimize (2.4). Hence it is very expensive in computational cost.

In the recent literature, Tibshirani (1996) proposed the LASSO, which is the solution of the penalized likelihood with L_1 penalty, defined by $p_\lambda(|\beta|) = \lambda|\beta|$. Furthermore, the L_q-penalty ($0 < q < 1$) yields a bridge regression (Frank and Friedman, 1993). The issue of selection of penalty function has been studied in depth by various authors (e.g. Antoniadis and Fan 2001). Fan and Li (2001) advocated that a good penalty function should yield an estimator with the following three properties: *unbiasedness* for a large true coefficient to avoid excessive estimation bias, *sparsity* (estimating a small coefficient as zero) to reduce model complexity, and *continuity* to avoid unnecessary variation in model prediction. However, none of L_q-penalties ($q \geq 0$) are satisfied with the necessary conditions for the above three properties (Fan and Li, 2001).

A simple penalty function that satisfies all the three mathematical requirements is the smoothly clipped absolute deviation (SCAD) penalty, defined by

$$p'_\lambda(\beta) = I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \text{ for some } a > 2 \quad \text{and } \beta > 0, \tag{3.1}$$

with $p_\lambda(0) = 0$. This penalty function involves two unknown parameters λ and a. Justifying from a Bayesian statistical point of view, Fan and Li (2001) suggested using $a = 3.7$. The Bayes risk cannot be reduced much with other choices of a and simultaneous selection of a and λ does not have any significant improvements from our experience. The SCAD improves the entropy penalty function in two aspects: saving computational cost and resulting in a continuous solution to avoid unnecessary modeling variation. Furthermore, the SCAD improves the L_1 penalty by avoiding excessive estimation bias, and improves the bridge regression by reducing modeling variation in model prediction.

Local Quadratic Approximation

Since the penalty functions such as the SCAD or the L_1 penalty are singular at the origin and may be nonconcave functions, it is challenging to find the solution of the penalized pseudo-partial likelihood function $\mathcal{L}(\boldsymbol{\beta})$ defined in (2.5). We will use a local quadratic approximation to the SCAD penalty. Suppose that we are given an initial value $\boldsymbol{\beta}^{(0)}$ that is close to the true value of $\boldsymbol{\beta}$. If $\beta_j^{(0)}$

is not close to 0, then the penalty function is locally approximated by a quadratic function as

$$p_{\lambda_j}(|\beta_j|) \approx q_{\lambda_j}(|\beta_j|) \equiv p_{\lambda_j}(|\beta_j^{(0)}|) + \frac{1}{2}\{p'_{\lambda_j}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)2}).$$

Otherwise, set $\widehat{\beta}_j = 0$. In our algorithm, we set $\widehat{\beta}_j = 0$ if its absolute value is less than λ_j. The above quadratic approximation has the property that

$$p_{\lambda_j}(|\beta_j^{(0)}|) = q_{\lambda_j}(|\beta_j^{(0)}|) \quad \text{and} \quad p'_{\lambda_j}(|\beta_j^{(0)}|) = q'_{\lambda_j}(|\beta_j^{(0)}|).$$

With the aid of the local quadratic approximation, a Newton-Raphson algorithm can be applied for the penalized pseudo-partial likelihood function with updating the approximation at each step during the iteration. In practice, we set the maximum pseudo-partial likelihood estimate $\widehat{\beta}^u$, the maximizer for $l(\beta)$ in (2.4), as the initial value of $\boldsymbol{\beta}$ since $\widehat{\beta}^u$ is root n consistent (see, for example, Clegg, Cai and Sen, 1999) and close to the true value. Moreover, the sandwich formula in Section 2.2 can be modified as follows:

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}) = \{\mathcal{L}''_a(\widehat{\boldsymbol{\beta}})\}^{-1}\widehat{\mathrm{cov}}\{\mathcal{L}'_a(\widehat{\boldsymbol{\beta}})\}\{\mathcal{L}''_a(\widehat{\boldsymbol{\beta}})\}^{-1},$$

where $\mathcal{L}_a(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^d q_{\lambda_j}(|\beta_j|)$. Therefore, $\mathcal{L}''_a(\widehat{\boldsymbol{\beta}}) = \ell''(\widehat{\boldsymbol{\beta}}) - n\Sigma_\lambda(\widehat{\boldsymbol{\beta}})$, where $\Sigma_\lambda(\widehat{\boldsymbol{\beta}}) = \mathrm{diag}\{p'_\lambda(|\widehat{\beta}_1|)/|\widehat{\beta}_1|, \cdots, p'_\lambda(|\widehat{\beta}_d|)/|\widehat{\beta}_d|\}$, and $\mathrm{cov}\{\mathcal{L}'_a(\widehat{\boldsymbol{\beta}})\}$ is estimated by $\widehat{\mathrm{cov}}\{\ell'(\widehat{\boldsymbol{\beta}})\}$. The sandwich formula applies only to nonzero estimated coefficients. The performance of this estimator will be examined in our simulation studies.

Selection of Tuning Parameter

We suggest to select the tuning parameter λ using data-driven approaches. In particular, we use the idea of the generalized cross validation (GCV, Graven and Wahba, 1977) to choose λ in our analysis and simulations. Note that the effective number of parameters for the penalized pseudo-partial likelihood in the last step of the Newton-Raphson algorithm iteration is

$$e(\lambda_1, \cdots, \lambda_d) = \mathrm{tr}[\{\mathcal{L}''_a(\widehat{\boldsymbol{\beta}})\}^{-1}\ell''(\widehat{\boldsymbol{\beta}})].$$

The generalized cross-validation statistic is defined by

$$\mathrm{GCV}(\lambda_1, \cdots, \lambda_d) = \frac{-\ell(\widehat{\boldsymbol{\beta}})}{n\{1 - e(\lambda_1, \cdots, \lambda_d)/n\}^2}$$

and $(\widehat{\lambda}_1, \cdots, \widehat{\lambda}_d)$ is the selected tuning parameter which minimizes the function $\mathrm{GCV}(\lambda_1, \cdots, \lambda_d)$. The minimization problem over a d-dimensional space is difficult. Intuitively, it is expected that the

magnitude of λ_j should be proportional to the standard error of the unpenalized maximum pseudo-partial likelihood estimator of β_j. In practice, we suggest taking $\lambda_j = \lambda \cdot SE(\widehat{\beta}_j^u)$, where $\widehat{\beta}_j^u$ is the maximum pseudo-partial likelihood estimator from maximizing (2.4) and $SE(\widehat{\beta}_j^u)$ is its standard error. Such a choice for λ_j works well from our simulation experience. Thus, the minimization problem will reduce to a one-dimensional problem, and the tuning parameter can be estimated by searching for the solution over a set of grid points.

Computing Algorithm for the Proposed Procedures

We summarize the above procedures in the following algorithm.

1. Choose a grid point set for λ, say, $(\lambda_1^*, \cdots, \lambda_S^*)$ and let $i = 1$.
2. With λ_i^*, compute the $\widehat{\lambda}_j = \lambda_i^* \cdot SE(\widehat{\beta}_j^u)$ $(j = 1, \cdots, d)$, where $\widehat{\beta}_j^u$ and $SE(\widehat{\beta}_j^u)$ are estimated from (2.4).
3. Plug $\widehat{\lambda}_j, j = 1, \cdots, d$, into (2.5) and use the Newton-Raphson algorithm to compute $\widehat{\beta}$.
4. Compute GCV_i. Let $i = i + 1$, and back to step 2. Repeat steps 2-4 until all S grid points are exhausted.
5. The final estimator for $\boldsymbol{\beta}$ is the one that has the lowest GCV.

4 Simulation Studies

In this section, extensive Monte Carlo simulations were conducted to assess the finite sample performance of the proposed penalized pseudo-partial likelihood procedure with the SCAD penalty. A naive approach is to delete insignificant variables using the 2-σ rule: delete a variable when its t-value is less than 2. We will compare the proposed procedure with this naive approach and the oracle procedure in terms of model error and model complexity. We further test the accuracy of the proposed standard error formula. All simulations were conducted using MATLAB codes.

4.1 Prediction error and model error

For a general regression model,

$$Y = \mu(\mathbf{x}) + \varepsilon,$$

let $\widehat{\mu}(\mathbf{x})$ be a prediction procedure constructed using the available data. When the covariate $\mathbf{x}$ is random, the prediction error is defined as

$$\mathrm{PE}(\widehat{\mu}) = E\{Y - \widehat{\mu}(\mathbf{x})\}^2,$$

where the expectation is only taken with respect to the new observation $(\mathbf{x}, Y)$. The prediction error can be decomposed as

$$\mathrm{PE}(\widehat{\mu}) = E\mathrm{var}(Y|\mathbf{x}) + E\{E(Y|\mathbf{x}) - \widehat{\mu}(\mathbf{x})\}^2.$$

The first component is inherent due to stochastic errors. The second component is due to lack of fit to an underlying model. This component is called a model error and is denoted by $\mathrm{ME}(\widehat{\mu})$. For the marginal hazards model (2.1),

$$h_{ijk}(t|\mathbf{x}) = h_{0j}(t)\exp(\boldsymbol{\beta}^T\mathbf{x}),$$

it follows by some calculation that the conditional mean

$$\mu_j(\mathbf{x}) = E(T_{\cdot j}|\mathbf{x}) = \int_0^\infty t\, h_{0j}(t)\exp(\mathbf{x}^T\boldsymbol{\beta})\exp\{-\int_0^t h_{0j}(u)\exp(\mathbf{x}^T\boldsymbol{\beta})\,du\}\,dt. \tag{4.1}$$

In our simulations, we will take $h_{0j}(t) \equiv h_j$. Thus,

$$\mu_j(\mathbf{x}) = h_j^{-1}\exp(-\mathbf{x}^T\boldsymbol{\beta}).$$

Therefore, the model error equals to

$$\mathrm{ME}(\widehat{\mu}_j) = h_j^{-2}E\{\exp(-\mathbf{x}^T\widehat{\boldsymbol{\beta}}) - \exp(-\mathbf{x}^T\boldsymbol{\beta})\}^2$$

for the marginal proportional hazards model. Note that the estimate $\widehat{\boldsymbol{\beta}}$ is root n consistent. Therefore, by the Taylor expansion,

$$\mathrm{ME}(\widehat{\mu}_j) \approx h_j^{-2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\{E\mathbf{x}\mathbf{x}^T\exp(-2\boldsymbol{\beta}^T\mathbf{x})\}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

which will be referred to as approximate model error (AME).

4.2 Simulation study

In our simulations, we take $J = K = 2$, and the failure times $T_{i11}, T_{i12}, T_{i21}$ and T_{i22} for the ith cluster are generated from the multivariate Clayton-Oakes distribution (Clayton and Cuzick, 1985,

Oakes, 1989) with a marginal exponential distribution for the two types of failure and for the two subjects in a cluster

$$\begin{aligned} & P\{T_{i11} > t_{i11}, T_{i12} > t_{i12}, T_{i21} > t_{i21}, T_{i2} > t_{i22} | \mathbf{x}_{ijk}, j = 1, 2, k = 1, 2\} \\ = & \left[\sum_{j=1}^{2} \sum_{k=1}^{2} \exp\{t_{ijk} \lambda_{0j} \theta^{-1} \exp(\boldsymbol{\beta}^T \mathbf{x}_{ijk})\} - 3 \right]^{-\theta}, \end{aligned}$$

where $\boldsymbol{\beta}$ equals $(\log(2), 0, 0, 0, -\log(3), 0, 0, 0.5\log(3), 0, 0, 0, 0)^T$, which is a 12-dimension vector consisting of 3 nonzero components and 9 zeros components. In our simulation, $\lambda_{01} = 1$ and $\lambda_{02} = 5$. The covariate vector $\mathbf{x}_{ijk}$ has a normal distribution with marginally standard normal and the correlation between x_{ijkl} and $x_{ijkl'}$ being $\rho^{|l-l'|}$ and $\rho = 0.5$. Censoring times C_{ijk} are generated from uniform distribution over $(0, c)$. We took $c = 5$ or 1. The censoring rates are approximately 18% and 45%, respectively. For the multivariate Clayton-Oakes distribution, $\theta \to 0$ gives the maximal positive correlation of 1 between failure times and $\theta \to \infty$ gives independence. In our simulation, θ was chosen to be 0.25, 1.5 or 5, which corresponds to high, moderate or low positive dependence, respectively. The number of clusters was taken as $n = 50$ and 100.

Model errors and approximate model errors of the proposed procedures are compared to those of the maximum pseudo-partial likelihood estimates from the full model. The model errors are computed by 10,000 Monte Carlo simulations since there is no closed form for the expectation involved in the definition of ME. We found that the results for RME, the ratio of ME, of an underlying procedure to that of the maximum pseudo-partial likelihood estimate, are very close to those for relative approximate model errors (RAME). Thus, we present only the results of RAME. Following Tibshirani (1996), the median and median of absolute deviation (MAD) of RAME over 500 simulated data sets are summarized in Table 1. The average number of zero coefficients is also reported in Table 1, in which the column labeled "C" stands for the average restricted only to the true zero coefficients, while the column label "I" depicts the average of coefficients erroneously estimated as 0. In Tables 1 and 2, "Naive" stands for the 2-σ rule procedure, and "Oracle" for the oracle procedure. Table 1 shows that the RAME of the SCAD is very close to that of oracle estimator, which is consistent with the theoretic result in Theorem 2.2. Furthermore, the SCAD reduces the model complexity almost as effectively as the oracle procedure. Compared with the SCAD, the 2-σ rule has much greater RAME. This demonstrates that the SCAD is more efficient than the 2-σ procedure, which is also evidenced from Table 2.

In Table 2, we examine the accuracy of the proposed standard error formula. The sample standard deviation, denoted by SD, of the estimated coefficients for the 500 simulated data sets can be regarded as the true standard errors for those estimators except Monte Carlo errors. The average of the estimated standard errors, denoted by SE, for the 500 simulated data sets, and their standard deviation, denoted by std(SE), of the estimated standard errors gauge the overall performance of the standard error formula. In our simulations, the standard errors of estimated coefficients were set to be 0 if they were estimated as zero. Table 2 depicts only the SD, SE, std(SE) and the 90% and 95% coverage probabilities of β_1 for the sample size $n = 50$ and 100 when $c = 1$ and $\theta = 5$. The results for all other cases are similar. From Table 2, the sandwich formula gives us accurate estimates of standard errors and the coverage probabilities which are close to the nominal level.

5 Analysis of the Framingham Study Data Set

We illustrate the proposed variable selection procedures by an analysis of a data set collected in the Framingham Heart Study (FHS). The FHS was initiated in 1948, with 2336 men and 2873 women aged between 30 and 62 years at their baseline examination (Dawber, 1980). In this study, multiple failure outcomes, for instance, times to coronary heart disease (CHD) and cerebrovascular accident (CVA), were observed from the same individual. In addition, as the primary sampling unit was the family, failure times recorded are likely to be dependent for the individuals within a family.

For simplicity, we consider only time to obtain first evidence of CHD and of CVA, and analyze only data for participants in the FHS study who had an examination at age 44 or 45 and were disease-free at that examination. By disease-free we mean that there exists no history of hypertension or glucose intolerance and no previous experience of a CHD or CVA. The time origin is the time of the examination at which an individual participated in the study and the follow up information is up to year 1980. The risk factors of interest are: body mass index (BMI), denoted by x_1, cholesterol level (x_2), systolic blood pressure (x_3), smoking status (x_4), coded by 1 if this individual is smoking, and 0 otherwise, gender (x_5), coded by 1 for female and 0 for male. The values of risk factors were taken from the biennial examination at which an individual was included in the sample. Because some individuals were in the study several years prior to inclusion into the data set, the waiting time, denoted by x_6, from entering the study to reaching 44 or 45 years of

age was used as a covariate to account for the cohort effect. Since x_1, x_2, x_3 and x_6 are continuous covariates, they are standardized individually in our analysis.

To explore possible nonlinear effects and interaction effects of the risk factors, we include all main effects, quadratic effects and interaction effects of the risk factors and covariates in the full model. This results in a mixed baseline hazard model with 50 covariates:

$$h_{ijk}(t, \mathbf{x}_{ijk}) = h_{0j}(t) \exp\{\boldsymbol{\beta}_j^T \mathbf{x}_{ijk}\}, \tag{5.1}$$

where $\mathbf{x}_{ijk}$ consists of all possible linear, quadratic and interaction terms of the risk factors and covariates x_1 to x_6. Model (5.1) allows different baseline hazards and different regression coefficients for CHD and CVA, but an identical baseline hazards for siblings.

The maximum pseudo-partial likelihood estimate for $\boldsymbol{\beta}$ is computed. The natural logarithm of the pseudo-partial likelihood for the full model of 50 covariates is -2017.9590. We then applied the naive approach (2-σ rule) to the full model (5.1). In the resulting model, only the linear main effect of *systolic blood pressure* (x_3) is significant for CHD, and there is no significant risk factor for CVA. The natural logarithm of the pseudo-partial likelihood for the selected model by the naive approach is -2067.9130.

Next we apply the SCAD procedure to model (5.1) to select significant variables. In the implementation of the SCAD procedure, since all covariates are important confounding variables, we included them in the model by not penalizing the linear main effect of x_1 to x_6. Thus, all linear effects are included in the selected model. The GCV method is used to select the regularization parameter. Figure 1 depicts the plot of GCV score versus λ. The regularization parameter λ equals 0.9053, selected by minimizing the GCV scores. The logarithm of the pseudo-partial likelihood for the model selected by the SCAD with the selected λ is -2022.6635. This represents an increase of the logarithm of the pseudo-partial likelihood by 10.1923 from that of the full model, which is much less than 25, the number of covariates excluded from the full model. From extension of Theorem 3 of Cai (1999), the limiting distribution of the pseudo-partial likelihood ratio statistic is a weighted sum of Chi-square distributions with 1 degree of freedom. Based on 100,000 Monte Carlo simulations, we computed the p-value, which equals 0.9926. This is in favor of the selected model. We further compared the selected model by SCAD with the one selected by the naive approach. The corresponding pseudo-partial likelihood ratio statistic is 90.4989 with p-value 0.0000 obtained by 100,000 Monte Carlo simulations. This is also in favor of the selected model by SCAD.

In another confirmation of the selected model, we compare the selected model with the linear main effect model which include only all the linear main effects of x_1 to x_6. The maximum pseudo-partial likelihood estimate for the unknown regression coefficients is computed, and the natural logarithm of the pseudo-partial likelihood for the linear main effect model is -2034.6527. The pseudo-partial likelihood ratio statistic for testing H_0: the linear main effect model versus H_1: the selected model, is 23.9783. Based on 100,000 Monte Carlo simulations, the corresponding p-value equals 0.0353. This indicates that the selected model fits the data better than the model with only the linear main effects.

The resulting estimate and standard error for $\boldsymbol{\beta}$ in the selected model is depicted in Table 3. For all terms associated with x_1, x_2, x_3 and x_6, the results in Table 3 are based on the standardized variables rather than the original ones. Table 3 clearly indicates that there are a few possible quadratic effects and many interactions among the risk factors on CHD and CVA. It shows that people with higher cholesterol level have higher risk in developing CHD. Due to the presence of the interaction between cholesterol level and smoking status, the hazard ratio is exp(0.0576+0.1550)=1.24 for smokers and exp(0.0576)=1.059 for nonsmokers, for an increase of 3.6 mg/dL (one standard deviation) in cholesterol level. For a given cholesterol level x_2, the hazard ratio for smokers relative to nonsmokers can be computed as $\exp(0.4754+0.1550x_2)$.

6 Conclusional Remarks

Stepwise deletion and subset selection are practically useful, but they ignore stochastic errors inherited in the stage of variable selections. Hence their theoretic properties are somewhat difficult to understand. Furthermore, as analyzed by Breiman (1996), the best subset variable selection suffers from several drawbacks, among which the most severe one is its lack of stability. We propose a class of variable-selection procedures via nonconcave penalized pseudo-partial likelihood for the marginal mixed baseline hazards model with multivariate failure time data. The selected model based on the nonconcave penalized pseudo-partial likelihood satisfies $p(|\beta_j|) = 0$ for certain β_j's. Therefore, model estimation is performed at the same time with model selection. Because we select variables and estimate parameters simultaneously, we can establish the rate of convergence and show that with properly chosen penalty functions some estimators will possess an oracle property. Our theoretic and numerical results show that the proposed procedure performs as powerfully as

the oracle estimator and is practical to use for model-selection in real applications.

A nice feature of the proposed method is that it allows the prior knowledge of subject-matter considerations to be incorporated into the model selection process while still provides a unified approach. This is demonstrated in the analysis of FHS study where we kept the six important confounding variables in the model by not penalizing their coefficients. Compared with other available model-selection criteria, such as the 2-σ method or the pseudo-partial likelihood ratio test (Cai, 1999), our method is a one-step, top-down approach in the sense that it selects a final model from a fully saturated model. Our final model includes some interactions, and quadratic terms that will not be picked up by the 2-σ method. Furthermore, the generalized pseudo-partial likelihood test (Cai 1999) confirmed that this final model fits the data better than the one with only the linear terms of the *a priori* covariate set. Considerable more work will be needed if one uses a step-wise approach like the GLR method to expand the model with only the linear effects to the final model.

The established rate of convergence and the oracle property of the penalized pseudo-partial likelihood estimator for the marginal model (2.1) can be easily extended to other marginal hazards models, such as (2.2), with a slightly different pseudo-partial likelihood function. For example, for the common baseline hazards model (2.2), we can use the following pseudo-partial likelihood:

$$\ell_c(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} \delta_{ijk}\left(\boldsymbol{\beta}^T\mathbf{x}_{ijk}(Z_{ijk}) - \log\left[\sum_{l=1}^{n}\sum_{m=1}^{J}\sum_{g=1}^{K} Y_{lmg}(Z_{ijk})\exp\{\boldsymbol{\beta}^T\mathbf{x}_{lmg}(Z_{ijk})\}\right]\right).$$

The corresponding asymptotic results in Theorems 2.1 and 2.2 for the estimator based on the penalized pseudo-partial likelihood $\ell_c(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|)$ can be established using similar arguments in Appendix.

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, **60**, 255-265.

Antoniadis, A. and Fan, J. (2001). Regularization wavelet approximations (with discussion). *J. Amer. Statist. Assoc.*, **96**, 939-967.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350-2383.

Cai, J. (1999). Hypothesis testing of hazard ratio parameters in marginal models for multivariate failure time data. *Lifetime Data Analysis*, **5**, 39-53.

Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, **82**, 151-164.

Cai, J. and Prentice, R. L. (1997). Regression estimation using multivariate time data and a common baseline hazard function model. *Lifetime Data Analysis*, **3**, 197-213.

Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Jour. Roy. Statist. Soc. Ser. A*, **148**, 82-117.

Clegg, L. X. , Cai, J. and Sen, P. K. (1999). A marginal mixed baseline hazards model for multivariate failure time data. *Biometrics*, **55**, 805-812.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Jour. Roy. Statist. Soc. Ser. B*, **34**, 187-220.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.

Dawber, T. R. (1980). *The Framingham Study, The Epidemiology of Atherosclerotic Disease*, Cambridge, MA, Harvard University Press.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Jour. Ameri. 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.

Faraggi, D. and Simon, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics*, **54**, 1475-1485.

Frank, I.E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-148.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data.* Springer, New York.

Lee, E. W., Wei, L. J. and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J. P. Klein and P. Goel, eds, Boston: Kluwer Academic Publishers, 237-248.

Liang, K.-Y., Self, S. G. and Chang, Y.-C. (1993). Modelling marginal hazards in multivariate failure time data. *J. Royal Statist. Soc., Ser. B*, **55**, 441-453.

Lin, D. Y. (1994), Cox regression analysis of multivariate failure time data: The marginal approach. *Statist. in Med.*, **13**, 2233-2247.

Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). *Jour. Roy. Statist. Soc., B*, **30**, 31-66.

Oakes, D. (1989). Bivariate survival models induced by frailty. *Jour. Amer. Statist. Assoc.*, **84**, 487-493.

Pötscher, B. M. (1989). Model selection under nonstationary: autoregressive model and stochastic linear regression models, *Ann. Statist.*, **17**, 1257-1274.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem, *Biometrika*, **76**, 369-374.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

Shao, J. (1997). An asymptotic theory for linear model selection, *Statistica Sinica*, **7**, 221-264.

Shao, J. and Rao, J. (2000). The GIC for model selection: a hypothesis testing approach. *J. Statist. Planning and Inference*, **77**, 103-117.

Spiekerman, C. F. and Lin, D. Y. (1998). Marginal regression models for multivariate failure time data, *Jour. Amer. Statist. Assoc.*, **93**, 1164-1175.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Jour. Roy. Statist. Soc., B*, **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statist. Med.*, **16**, 385-395.

Wei, L. J., Lin, D. Y. and Weisseld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.*, **84**, 1065-1073.

Appendix

To facilitate the notation, let $N_{ijk}(t) = I(Z_{ijk} \leq t, \delta_{ijk} = 1)$ be the counting process, and $h_{ijk}(t)$ and $M_{ijk}(t) = N_{ijk}(t) - \int_0^t Y_{ijk}(u)h_{ijk}(u)\,du$ be their corresponding marginal hazards function and marginal martingale, respectively, with respect to the filtration $\mathcal{F}_{jk}(t^-)$, where $\mathcal{F}_{jk}(t)$ is the σ-field generated by $\{N_{ijk}(u), Y_{i11}(u), ..., Y_{iJK}(u), \mathbf{x}_{i11}(u), ..., \mathbf{x}_{iJK}(u);\ 0 \leq u \leq t, i = 1, ..., n\}$. Define

$$\begin{aligned}
\mathbf{S}_{jk}^{(d)}(\boldsymbol{\beta}; t) &= \frac{1}{n}\sum_{i=1}^{n} Y_{ijk}(t)\mathbf{x}_{ijk}(t)^{\otimes d}\exp\{\boldsymbol{\beta}^T\mathbf{x}_{ijk}(t)\}, \quad d = 0, 1, 2 \\
\mathbf{S}_{j\cdot}^{(d)}(\boldsymbol{\beta}; t) &= \sum_{k=1}^{K}\mathbf{S}_{jk}^{(d)}(\boldsymbol{\beta}; t), \quad d = 0, 1, 2, \\
\mathbf{E}_j(\boldsymbol{\beta}; t) &= \mathbf{S}_{j\cdot}^{(1)}(\boldsymbol{\beta}; t)/\mathbf{S}_{j\cdot}^{(0)}(\boldsymbol{\beta}; t), \\
\mathbf{V}_j(\boldsymbol{\beta}; t) &= \mathbf{S}_{j\cdot}^{(2)}(\boldsymbol{\beta}, t)/\mathbf{S}_{j\cdot}^{(0)}(\boldsymbol{\beta}, t) - \mathbf{E}_j(\boldsymbol{\beta}; t)^{\otimes 2},
\end{aligned}$$

where $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = 1$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for a vector $\mathbf{a}$.

Regularity conditions:

(A) For simplicity, assume that T_{ijk} takes values on a finite interval $[0, \tau]$, and $\int_0^\tau h_{0j}(t)\, dt < \infty$ for $j = 1, \cdots, J$.

(B) There exists a neighborhood $\mathcal{B}$ of the true value $\boldsymbol{\beta}_0$ that satisfies each of the following conditions: (1) there exists a scalar, vector, and matrix function $\mathbf{s}_{jk}^{(d)}(\boldsymbol{\beta}, t)$ $(d = 0, 1, 2)$ defined on $\mathcal{B} \times [0, \tau]$ such that $\sup_{t\in[0,\tau],\boldsymbol{\beta}\in\mathcal{B}} \|\mathbf{S}_{jk}^{(d)}(\boldsymbol{\beta}, t) - \mathbf{s}_{jk}^{(d)}(\boldsymbol{\beta}, t)\| \to 0$ in probability; (2) there exists a matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\beta})$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{var}(\mathbf{D}_i) \to \mathbf{D},$$

where

$$\mathbf{D}_i = \sum_{j=1}^{J}\sum_{k=1}^{K} \int_0^\tau \{\mathbf{x}_{ijg}(t) - \mathbf{e}_j(\boldsymbol{\beta}_0; t)\}\, dM_{ijk}(t),$$

and $\mathbf{e}_j(\boldsymbol{\beta}; t) = \left\{\sum_{k=1}^{K} \mathbf{s}_{jk}^{(2)}(\boldsymbol{\beta}; t)\right\} / \left\{\sum_{k=1}^{K} \mathbf{s}_{jk}^{(0)}(\beta; t)\right\}$.

(C) Let $\mathbf{s}_{jk}^{(d)}$, $d = 0, 1, 2$, $\mathcal{B}$ and $\mathbf{e}_j$ be as in Condition (B) and define $\mathbf{v}_j = \left\{\sum_{k=1}^{K} \mathbf{s}_{jk}^{(2)}(\boldsymbol{\beta}, t)\right\} / \left\{\sum_{k=1}^{K} \mathbf{s}_{jk}^{(0)}(\boldsymbol{\beta}, t)\right\} - \mathbf{e}_j(\boldsymbol{\beta}; t)^{\otimes 2}$. Then for all $\boldsymbol{\beta} \in \mathcal{B}$, $t \in [0, \tau]$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$: $\mathbf{s}_{jk}^{(1)}(\boldsymbol{\beta}; t) = \partial \mathbf{s}_{jk}(\boldsymbol{\beta}; t)/\partial\boldsymbol{\beta}$ and $\mathbf{s}_{jk}^{(2)}(\boldsymbol{\beta}; t) = \partial \mathbf{s}_{jk}^{(1)}(\boldsymbol{\beta}; t)/\partial\boldsymbol{\beta}$. Assume $\mathbf{s}_{jk}^{(0)}(\boldsymbol{\beta}; t)$ is bounded away from 0 on $\mathcal{B}\times[0, \tau]$, and the matrix $\sum_{j=1}^{J} \int_0^\tau \mathbf{v}_j(\boldsymbol{\beta}_0; t) \sum_{k=1}^{K} \mathbf{s}_{jk}^{(0)}(\boldsymbol{\beta}_0; t) h_{0j}(t)\, dt$ is positive definite.

(D) In probability

$$\frac{1}{n}\sum_{i=1}^{n} E\{\|\mathbf{D}_i\|^2 I(\|\mathbf{D}_i\| > \varepsilon n^{1/2})\} \to 0.$$

These conditions are adapted from Clegg, Cai and Sen (1999) and guarantee the asymptotic normality of the pseudo-partial likelihood estimator, the maximizer of $\ell(\boldsymbol{\beta})$ defined in (2.4). Under these conditions, there exists a sequence $\boldsymbol{\beta}_n \to \boldsymbol{\beta}_0$ as $n \to \infty$.

Proof of Theorem 2.1. Denote by

$$\ell''(\boldsymbol{\beta}) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}.$$

In term of counting process,

$$\ell''(\boldsymbol{\beta}) = -\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} \int_0^\tau \mathbf{V}_j(\boldsymbol{\beta}, u)\, dN_{ijk}(u).$$

Under Conditions (A), (B) and (D),

$$n^{-1}\ell''(\boldsymbol{\beta}) \to -A(\boldsymbol{\beta}),$$

in probability, uniformly in $\boldsymbol{\beta} \in \mathcal{B}$, where

$$A(\boldsymbol{\beta}) = \sum_{j=1}^{J} \int_0^\tau \mathbf{v}_j(\boldsymbol{\beta}; t) \sum_{k=1}^{K} \{\mathbf{s}_{jk}^{(0)}(\boldsymbol{\beta}; t)\} h_{0j}(t)\, dt.$$

The matrix $A(\boldsymbol{\beta})$ is a finite positive definite matrix. Thus, using Taylor's expansion, for any $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$, it follows that

$$\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}_0) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \ell'(\boldsymbol{\beta}_0) - \frac{n}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T A(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\{1 + o_P(1)\}. \tag{A.1}$$

Let $\alpha_n = n^{-1/2} + a_n$. To prove Theorem 2.1, it is sufficient to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{ \sup_{\|\mathbf{u}\|=C} \mathcal{L}(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) < \mathcal{L}(\boldsymbol{\beta}_0) \right\} \geq 1 - \varepsilon. \tag{A.2}$$

This implies that with probability at least $1 - \varepsilon$ that there exists a local maximum in the ball $\{\boldsymbol{\beta}_0 + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$. Hence, there exists a local maximizer such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\alpha_n)$.

Using (A.1),

$$\ell(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\beta}_0) = \alpha_n \mathbf{u}^T \ell'(\boldsymbol{\beta}_0) - \frac{n\alpha_n^2}{2} \mathbf{u}^T A(\boldsymbol{\beta}_0)\mathbf{u}\{1 + o_P(1)\}.$$

Note that $p_\lambda(0) = 0$ and $p_\lambda(\cdot) \geq 0$, we have

$$\begin{aligned}
& \mathcal{L}(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \mathcal{L}(\boldsymbol{\beta}_0) \\
\leq\ & \ell(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\beta}_0) - n \sum_{j=1}^{s} \{p_{\lambda_{jn}}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_{jn}}(|\beta_{j0}|)\} \\
=\ & \alpha_n \mathbf{u}^T \ell'(\boldsymbol{\beta}_0) - \frac{n\alpha_n^2}{2} \mathbf{u}^T A(\boldsymbol{\beta}_0)\mathbf{u}\{1 + o_P(1)\} \\
& -n \sum_{j=1}^{s} \{p_{\lambda_{jn}}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_{jn}}(|\beta_{j0}|)\}
\end{aligned} \tag{A.3}$$

Note that $n^{-1/2}\ell'(\boldsymbol{\beta}_0) = O_P(1)$ (Clegg, Cai and Sen, 1999). Thus, the first term on the right-hand side of (A.3) is on the order $O_P(n^{1/2}\alpha_n) = O_P(n\alpha_n^2)$. By choosing a sufficiently large C, the second term dominates the first term uniformly in $\|\mathbf{u}\| = C$. Note that, by applying Taylor expansion and Cauchy-Schwarz inequality, the third term in (A.3) is bounded by

$$\sqrt{s} n \alpha_n a_n \|\mathbf{u}\| + n\alpha_n^2 b_n \|\mathbf{u}\|^2.$$

This is also dominated by the second term of (A.3). Hence, by choosing a sufficiently large C, (A.2) holds. This completes the proof.

Proof of Theorem 2.2.

To establish the oracle property, we first prove that the resulting estimator of the penalized pseudo-partial likelihood estimator must possess the sparsity property: $\widehat{\boldsymbol{\beta}}_2 = 0$. To this end, we show that with probability tending to 1, for any given $\boldsymbol{\beta}_1$ satisfying that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and any constant C, it holds that

$$\mathcal{L}\{(\boldsymbol{\beta}_1^T, \mathbf{0})^T\} = \max_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} \mathcal{L}\{(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T\}. \tag{A.4}$$

It is sufficient to show that for any $\boldsymbol{\beta}_1$ satisfying that $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_P(n^{-1/2})$, and $\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}$, $\partial\mathcal{L}(\boldsymbol{\beta})/\partial\beta_j$ and β_j have different signs for $\beta_j \in (-Cn^{-1/2}, Cn^{-1/2})$ for $j = s+1, \cdots, d$. By the definition of $\mathcal{L}(\boldsymbol{\beta})$,

$$\frac{\partial\mathcal{L}(\boldsymbol{\beta})}{\partial\beta_j} = \frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_j} - np'_{\lambda_{jn}}(|\beta_j|)\mathrm{sgn}(\beta_j).$$

Using Taylor's expansion, it follows that

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_j} = \frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\beta_j} - n\sum_{l=1}^{d} A_{jl}(\boldsymbol{\beta}_0)(\beta_l - \beta_{l0})\{1 + o_P(1)\},$$

where $A_{jl}(\boldsymbol{\beta}_0)$ is the (j,l)-element of $A(\boldsymbol{\beta}_0)$. By the assumption that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$ and the fact that $\partial\ell(\boldsymbol{\beta}_0)/\partial\beta_j = O_P(n^{1/2})$, it holds that

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_j} = O_P(n^{1/2}).$$

Therefore,

$$\frac{\partial\mathcal{L}(\boldsymbol{\beta})}{\partial\beta_j} = n\lambda_{jn}\left\{-\lambda_{jn}^{-1}p'_{\lambda_{jn}}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_P(\frac{1}{\sqrt{n}\lambda_{jn}})\right\}.$$

Since $O_P\{1/(\sqrt{n}\lambda_j)\} = o_P(1)$ by the assumption that $\sqrt{n}\lambda_j \to \infty$, the sign of $\partial\mathcal{L}(\boldsymbol{\beta})/\partial\beta_j$ is determined by that of β_j. This completes the proof of (A.4). Therefore, the penalized pseudo-partial likelihood estimator must possess the sparsity property: $\widehat{\boldsymbol{\beta}}_2 = 0$.

Next we prove the asymptotic normality of $\widehat{\boldsymbol{\beta}}_1$. It can be shown that there exists a $\widehat{\boldsymbol{\beta}}_1$ in Theorem 1 that is a root-n consistent local maximizer of $\mathcal{L}\{(\boldsymbol{\beta}_1, \mathbf{0})^T\}$ and that satisfies the penalized pseudo-partial likelihood equations

$$\partial\mathcal{L}\{(\widehat{\boldsymbol{\beta}}_1, \mathbf{0})^T\}/\partial\beta_j = 0$$

for $j = 1, \cdots, s$. Note that $\widehat{\boldsymbol{\beta}}_1$ is a consistent estimator. Using the Taylor expansion to $\partial\mathcal{L}\{(\widehat{\boldsymbol{\beta}}_1, \mathbf{0})^T\}/\partial\beta_j$ at $\boldsymbol{\beta}_0$, and the asymptotic normality of $\ell'(\boldsymbol{\beta}_0)$ (Clegg, Cai and Sen, 1999), the asymptotic normality of $\boldsymbol{\beta}_1$ in (2.9) can be established by Slutsky's theorem.

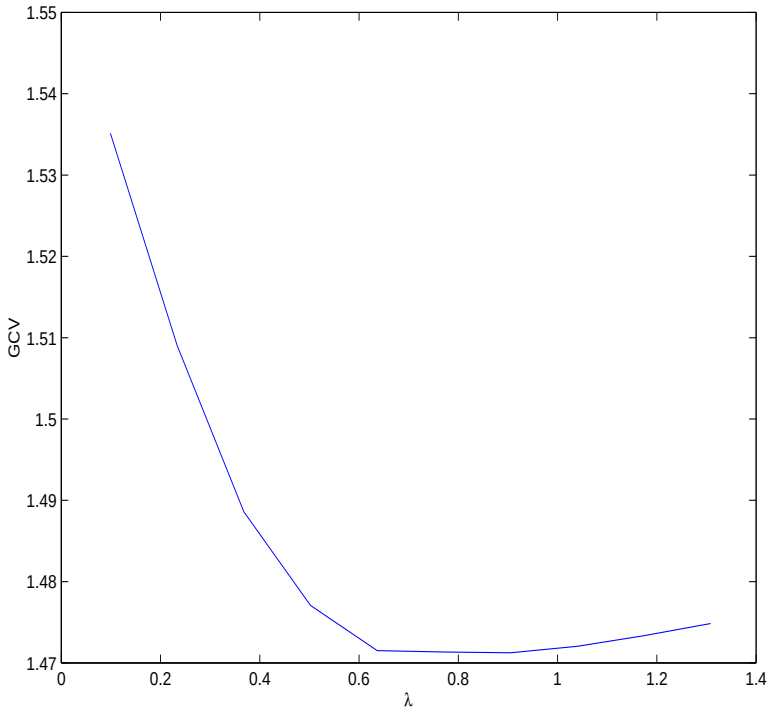

Figure 1: *Plot of Generalized Cross-Validation for the Framingham Heart Study Analysis*

Table 1: Relative Approximate Model Errors

		$c = 1$			$c = 5$		
θ	Method	RAME	Zero Coefficient		RAME	Zero Coefficient	
			$n = 50$				
		median (MAD)	C	I	median (MAD)	C	I
0.25	Naive	0.825(0.327)	8.562	0.024	0.836(0.292)	8.554	0.002
	SCAD	0.524(0.253)	8.860	0.036	0.560(0.262)	8.896	0.006
	Oracle	0.469(0.249)	9	0	0.525(0.267)	9	0
1.5	Naive	0.831(0.311)	8.530	0.030	0.792(0.317)	8.578	0
	SCAD	0.458(0.242)	8.848	0.036	0.507(0.261)	8.882	0
	Oracle	0.388(0.225)	9	0	0.471(0.244)	9	0
5	Naive	0.822(0.304)	8.516	0.026	0.867(0.327)	8.540	0.008
	SCAD	0.444(0.244)	8.852	0.040	0.433(0.224)	8.862	0.008
	Oracle	0.356(0.214)	9	0	0.374(0.216)	9	0
			$n = 100$				
0.25	Naive	0.842(0.315)	8.558	0	0.847(0.288)	8.600	0
	SCAD	0.547(0.229)	8.926	0.002	0.637(0.245)	8.944	0
	Oracle	0.523(0.235)	9	0	0.621(0.261)	9	0
1.5	Naive	0.846(0.292)	8.494	0	0.841(0.317)	8.634	0
	SCAD	0.526(0.250)	8.912	0	0.570(0.243)	8.958	0
	Oracle	0.480(0.254)	9	0	0.554(0.243)	9	0
5	Naive	0.777(0.329)	8.548	0	0.806(0.291)	8.564	0
	SCAD	0.451(0.209)	8.940	0	0.505(0.238)	8.942	0
	Oracle	0.431(0.206)	9	0	0.477(0.250)	9	0

Table 2: Standard Deviations and Standard Errors of $\widehat{\beta}_1$ ($c = 1$ and $\theta = 5$)

n	Method	SD	SE (std(SE))	90% coverage	95% coverage
50	Naive	0.1384	0.1369 (0.0216)	0.926	0.974
	SCAD	0.1163	0.0988 (0.0134)	0.890	0.952
	Oracle	0.1136	0.1113 (0.0175)	0.902	0.958
100	Naive	0.0933	0.0903 (0.0094)	0.914	0.948
	SCAD	0.0787	0.0743 (0.0071)	0.900	0.952
	Oracle	0.0769	0.0764 (0.0079)	0.900	0.944

Table 3: Estimated Coefficients and Standard Errors for the FHS data

Effect	CHD $\widehat{\beta}$(SE($\widehat{\beta}$))	CVA $\widehat{\beta}$(SE($\widehat{\beta}$))
x_1	0.0810 (0.1288)	0.4773 (0.2423)
x_2	0.0576 (0.1200)	-0.2409 (0.2655)
x_3	0.4129 (0.1570)	0.2917 (0.1477)
x_4	0.4754 (0.2361)	0.7077 (0.3587)
x_5	-0.3666 (0.2543)	-0.1016 (0.2890)
x_6	0.0249 (0.0802)	-0.1395 (0.1916)
x_1^2	-0.0743 (0.0512)	0 (−)
x_2^2	0 (−)	-0.0768 (0.1052)
x_3^2	0 (−)	0 (−)
x_6^2	0 (−)	0.2062 (0.1229)
$x_1 * x_2$	0 (−)	0 (−)
$x_1 * x_3$	0 (−)	-0.2224 (0.1435)
$x_1 * x_4$	0.1409 (0.1495)	-0.2207 (0.2628)
$x_1 * x_5$	0 (−)	0 (−)
$x_1 * x_6$	-0.1060 (0.0808)	0 (−)
$x_2 * x_3$	0 (−)	0 (−)
$x_2 * x_4$	0.1550 (0.1425)	0.5702 (0.3766)
$x_2 * x_5$	0 (−)	0 (−)
$x_2 * x_6$	0 (−)	0 (−)
$x_3 * x_4$	-0.1952 (0.1489)	0 (−)
$x_3 * x_5$	-0.2054 (0.1378)	0 (−)
$x_3 * x_6$	0 (−)	0 (−)
$x_4 * x_5$	-0.3071 (0.3106)	0 (−)
$x_4 * x_6$	0 (−)	0 (−)
$x_5 * x_6$	0 (−)	0.5753 (0.2545)