The Annals of Statistics
2002, Vol. 30, No. 1, 74–99

VARIABLE SELECTION FOR COX'S PROPORTIONAL HAZARDS MODEL AND FRAILTY MODEL

BY JIANQING FAN[1] AND RUNZE LI[2]

Chinese University of Hong Kong and Pennsylvania State University

A class of variable selection procedures for parametric models via nonconcave penalized likelihood was proposed in Fan and Li (2001a). It has been shown there that the resulting procedures perform as well as if the subset of significant variables were known in advance. Such a property is called an oracle property. The proposed procedures were illustrated in the context of linear regression, robust linear regression and generalized linear models. In this paper, the nonconcave penalized likelihood approach is extended further to the Cox proportional hazards model and the Cox proportional hazards frailty model, two commonly used semi-parametric models in survival analysis. As a result, new variable selection procedures for these two commonly-used models are proposed. It is demonstrated how the rates of convergence depend on the regularization parameter in the penalty function. Further, with a proper choice of the regularization parameter and the penalty function, the proposed estimators possess an oracle property. Standard error formulae are derived and their accuracies are empirically tested. Simulation studies show that the proposed procedures are more stable in prediction and more effective in computation than the best subset variable selection, and they reduce model complexity as effectively as the best subset variable selection. Compared with the LASSO, which is the penalized likelihood method with the L_1-penalty, proposed by Tibshirani, the newly proposed approaches have better theoretic properties and finite sample performance.

1. Introduction. An objective of survival analysis is to identify the risk factors and their risk contributions. Often, many covariates are collected and to reduce possible modeling bias, a large parametric model is built. An important and challenging task is to efficiently select a subset of significant variables upon which the hazard function depends. There are many variable selection techniques in linear regression models. Some of them have been extended to the context of censored survival data analysis, such as the best subset variable selection and stepwise deletion. Bayesian variable selection methods for censored survival data were proposed by Faraggi and Simon (1998), based on an idea of Lindley (1968).

Received November 2000; revised August 2001.

[1]Supported in part by NIH Grant IR01CA92571-01, NSF Grant DMS-01-96041 and RGC Grant CUHK4262/01P of HKSAR.

[2]Supported by NSF Grant DMS-01-02505.

AMS 2000 subject classifications. 62F12, 62N02.

Key words and phrases. Cox's regression model, frailty model, LASSO, penalized likelihood, partial likelihood, profile likelihood.

Despite their popularity, the sampling properties of the aforementioned selection methods are largely unknown and confidence intervals derived from the selected variables may not have right coverage probabilities.

Fan and Li (2001a) proposed a family of new variable selection methods based on a nonconcave penalized likelihood approach. The proposed methods are different from traditional approaches of variable selection in that they delete insignificant variables by estimating their coefficients as 0. Thus their approaches simultaneously select significant variables and estimate regression coefficients. LASSO, proposed by Tibshirani (1996, 1997), is a member of this family with the L_1-penalty. See also Knight and Fu (2000) for asymptotic properties of lasso-type estimators. The penalized likelihood approach was applied to linear regression, robust linear regression and generalized linear models. From their simulations, Fan and Li (2001a) showed the proposed penalized likelihood estimator with smoothly clipped absolute deviation penalty (defined in Section 2, the name of SCAD refers to the procedures related to this penalty function) outperforms the best subset variable selection in terms of computational cost and stability, in the terminology of Breiman (1996). The SCAD improves the LASSO via reducing estimation bias. Furthermore, they showed that the SCAD possesses an oracle property with a proper choice of regularization parameter, in the terminology of Donoho and Johnstone (1994). Namely, the true regression coefficients that are zero are automatically estimated as zero, and the remaining coefficients are estimated as well as if the correct submodel were known in advance. Hence, the SCAD and its siblings are an ideal procedure for variable selection, at least from the theoretical point of view. This encourages us to investigate their properties in Cox's proportional hazards model and frailty model, two popularly used semiparametric models.

It will be shown that the proposed penalized likelihood for the Cox regression model is equivalent to a penalized partial likelihood. This new approach can select significant variables and estimate regression coefficients simultaneously. This allows one to construct a confidence interval for coefficients easily. Rates of convergence of the penalized partial likelihood estimators are established. Further, with proper choice of regularization parameters, we will show that the SCAD performs as well as an oracle estimator. The significance of this is that the proposed procedure outperforms the maximum partial likelihood estimator when true coefficients have zero components and performs as well as if one knew the true submodel. This result is closely related to the super-efficiency phenomenon, given by the Hodges example [Lehmann (1983), page 405]. In addition, a modified Newton–Raphson algorithm is developed for maximizing the penalized partial likelihood function, and a standard error formula for estimated coefficients of nonzero components is derived by using a sandwich formula. The standard error formula is empirically tested for the Cox regression model. It performs very well with moderate sample sizes. The proposed method compares favorably with

the best subset variable selection, in terms of performance, model stability and computation.

Unlike the Cox regression model, there are some challenges in parameter estimation in the Cox frailty model even without the task of model selection. In fact, with the "least informative" nonparametric modeling for the baseline cumulative hazard function, the corresponding profile likelihood of the frailty model does not have a closed form. This poses some challenges to find estimates for parameters of interest. A new iterative procedure for this semi-parametric frailty model is proposed in order to find the profile maximum likelihood estimator. It provides a useful alternative to the EM algorithm for the frailty model, even without the task of variable selection. Standard error formulas are derived and empirically tested. Further, the penalized likelihood approach is extended to the semiparametric frailty model via penalizing the profile likelihood function. Due to its simultaneous selection of significant variables and estimation of regression coefficients, our approach allows one to construct confidence intervals for unknown coefficients via a sandwich formula. The corresponding sandwich formula is an estimator for the covariance matrix of the estimated coefficients. The Newton–Raphson algorithm with some modifications is used to find the solution of penalized profile likelihood score equations. It performs very well for moderate sample size. Again, in this model, SCAD outperforms the best subset variable selection and LASSO.

The paper is organized as follows. Motivations of variable selection via nonconcave penalized likelihood are briefly given in Section 2. A new variable selection procedure for the Cox model and Cox frailty model is proposed in Section 3, in which the consistency and an oracle property of the proposed procedures are established. A modification of the Newton–Raphson algorithm and standard error formulae for estimated coefficients are also presented in Section 3. Section 4 gives numerical comparisons among the newly proposed approach, the LASSO and the best subset variable selection. Proofs of main results are given in Section 5.

2. Variable selection via nonconcave penalized likelihood. Assume that the collected data $(\mathbf{x}_i, Y_i)$ are independent samples. Conditioning on $\mathbf{x}_i$, Y_i has a density $f_i(y_i; \mathbf{x}_i^T \boldsymbol{\beta})$. Denote by $\ell_i = \log f_i$, the conditional log-likelihood of Y_i given $\mathbf{x}_i$. As discussed in Fan and Li (2001a), a general form of penalized likelihood is

$$\sum_{i=1}^{n} \ell_i\big(y_i; \mathbf{x}_i^T \boldsymbol{\beta}\big) - n \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.1}$$

where d is the dimension of $\boldsymbol{\beta}$, $p_\lambda(\cdot)$ is a penalty function and λ is a tuning parameter (more generally, it is allowed to use λ_j). Some conditions on $p_\lambda(|\cdot|)$ are needed in order for the approach to be an effective variable selection procedure

[Antoniadis and Fan (2001)]. In particular, $p_\lambda(|\cdot|)$ should be irregular at the origin, that is, $p'_\lambda(0+) > 0$. Denote by $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$, and let $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{d0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$. Without loss of generality, it is assumed that $\boldsymbol{\beta}_{20} = \mathbf{0}$, and all components of $\boldsymbol{\beta}_{10}$ are not equal to 0. Under some regularity conditions, Fan and Li (2001a) showed their SCAD estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ possesses the following oracle property. With probability tending to 1, for certain choice of $p_{\lambda_n}(\cdot)$, we have $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \to N\{\mathbf{0}, I_1^{-1}(\boldsymbol{\beta}_{10}, \mathbf{0})\},$$

where $I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$ is the Fisher information matrix for $\boldsymbol{\beta}_1$ knowing $\boldsymbol{\beta}_2 = \mathbf{0}$.

For linear regression models, when the columns of the design matrix $\mathbf{X}$ are orthonormal, it is easy to show that the best subset selection and stepwise deletion are equivalent to the penalized least squares estimator with the hard thresholding penalty, defined by

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda).$$

This penalty function was proposed by Fan (1997) and improved by Antoniadis (1997). The name HARD refers to the procedure related to the hard thresholding penalty. The hard thresholding penalty does not overpenalize the large value of $|\theta|$. Note that when a design matrix is not orthonormal, the penalized least-squares, the stepwise deletion and the best subset methods may not be equivalent. Other penalty functions have been used in the literature. The L_2-penalty $p_\lambda(|\theta|) = \lambda|\theta|^2$ results in a ridge regression. The L_1-penalty $p_\lambda(|\theta|) = \lambda|\theta|$ yields LASSO, proposed by Donoho and Johnstone (1994) in the wavelet setting and extended by Tibshirani (1996, 1997) to general likelihood settings.

A good penalty function should result in 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. Necessary conditions for unbiasedness, sparsity and continuity have been derived by Antoniadis and Fan (2001). However, all of the L_1, L_2 (indeed all of L_p-penalty) and the HARD penalties do not simultaneously satisfy these three mathematical conditions.

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

$$p'_\lambda(\theta) = I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \qquad \text{for some } a > 2 \text{ and } \theta > 0. \tag{2.2}$$

This function was proposed by Fan (1997) and involves two unknown parameters λ and a. In practice, one could search the best pair (λ, a) over two dimensional grids using some criteria, such as cross-validation and generalized cross-validation

[Craven and Wahba (1979)]. However, such an implementation can be computationally expensive. From Bayesian statistical point of view, Fan and Li (2001a) suggested using $a = 3.7$ and this value will be used throughout the whole paper.

Figure 1 depicts the aforementioned penalty functions. From Figure 1, the L_1, HARD and SCAD penalties are irregular at the origin, satisfying $p'_\lambda(0+) > 0$. This is a necessary condition for the penalized least-squares to possess the sparsity condition solution [Antoniadis and Fan (1999)]. The HARD and SCAD penalties are constant when θ is large. This does not excessively penalize large coefficients. However, SCAD is smoother than HARD and hence yields a continuous estimator.

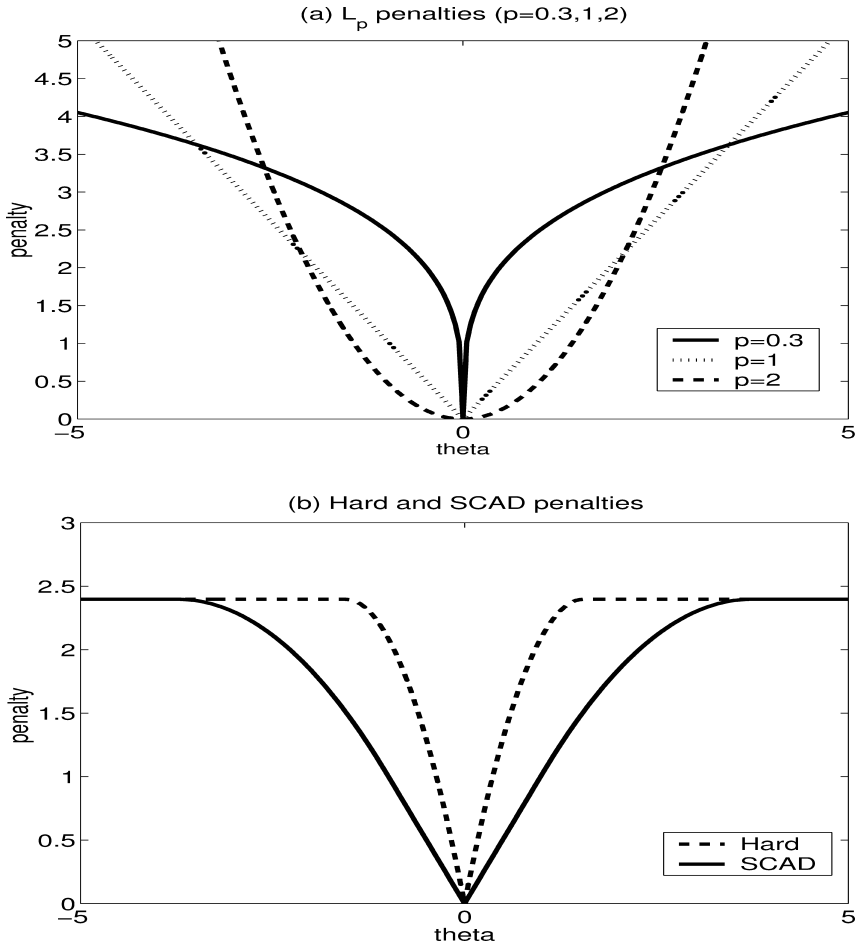

FIG. 1. *Plots of penalty functions.* (a) L_p *penalties, solid, dotted and dashed curves are for* $p = 0.3,\ 1$ *and* 2*, respectively.* (b) *Hard and SCAD penalties, solid and dashed curves are for the SCAD and hard penalties, respectively.*

3. Proportional hazards models. Let T, C and $\mathbf{x}$ be respectively the survival time, the censoring time and their associated covariates. Correspondingly, let $Z = \min\{T, C\}$ be the observed time and $\delta = I(T \le C)$ be the censoring indicator. It is assumed that T and C are conditionally independent given $\mathbf{x}$ and that the censoring mechanism is noninformative. When the observed data $\{(\mathbf{x}_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ is an independently and identically distributed random sample from a certain population $(\mathbf{x}, Z, \delta)$, a complete likelihood of the data is given by

$$L = \prod_u f(Z_i|\mathbf{x}_i) \prod_c \overline{F}(Z_i|\mathbf{x}_i) = \prod_u h(Z_i|\mathbf{x}_i) \prod_{i=1}^n \overline{F}(Z_i|\mathbf{x}_i), \tag{3.1}$$

where the subscripts c and u denote the product of the censored and uncensored data respectively, and $f(t|\mathbf{x})$, $\overline{F}(t|\mathbf{x})$ and $h(t|\mathbf{x})$ are the conditional density function, the conditional survival function and the conditional hazard function of T given $\mathbf{x}$. Statistical inference in this paper will be based on the likelihood function (3.1).

To present explicitly the likelihood function of Cox's proportional hazards model, more notation is needed. Let $t_1^0 < \cdots < t_N^0$ denote the ordered observed failure times. Let (j) provide the label for the item falling at t_j^0 so that the covariates associated with the N failures are $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(N)}$. Let R_j denote the risk set right before the time t_j^0:

$$R_j = \{i : Z_i \ge t_j^0\}.$$

Consider proportional hazards models,

$$h(t|\mathbf{x}) = h_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta}), \tag{3.2}$$

with the baseline hazard functions $h_0(t)$ and parameter $\boldsymbol{\beta}$. The likelihood in (3.1) becomes

$$L = \prod_{i=1}^N h_0(Z_{(i)}) \exp(\mathbf{x}_{(i)}^T \boldsymbol{\beta}) \prod_{i=1}^n \exp\{-H_0(Z_i) \exp(\mathbf{x}_i^T \boldsymbol{\beta})\},$$

where $H_0(\cdot)$ is the cumulative baseline hazard function. If the baseline hazard function has a parametric form, $h_0(\boldsymbol{\theta}, \cdot)$ say, then the corresponding penalized log-likelihood function is

$$\begin{aligned} &\sum_{i=1}^N [\log\{h_0(\boldsymbol{\theta}, Z_{(i)})\} + \mathbf{x}_{(i)}^T \boldsymbol{\beta}] \\ &\quad - \sum_{i=1}^n \{H_0(\boldsymbol{\theta}, Z_i) \exp(\mathbf{x}_i^T \boldsymbol{\beta})\} - n \sum_{j=1}^d p_\lambda(|\beta_j|). \end{aligned} \tag{3.3}$$

Maximizing (3.3) with respect to $(\boldsymbol{\theta}, \boldsymbol{\beta})$ yields the maximum penalized likelihood estimator.

3.1. *Penalized partial likelihood.* In the Cox proportional hazards model, the baseline hazard function is unknown and has not been parameterized. Following Breslow's idea, consider the "least informative" nonparametric modeling for $H_0(\cdot)$, in which $H_0(t)$ has a possible jump h_j at the observed failure time t_j^0. More precisely, let $H_0(t) = \sum_{j=1}^N h_j I(t_j^0 \le t)$. Then

$$H_0(Z_i) = \sum_{j=1}^{N} h_j I(i \in R_j). \tag{3.4}$$

Using (3.4), the logarithm of penalized likelihood function of (3.3) becomes

$$\begin{aligned} &\sum_{j=1}^{N} \{\log(h_j) + \mathbf{x}_{(j)}^T \boldsymbol{\beta}\} \\ &\quad - \sum_{i=1}^{n} \left\{ \sum_{j=1}^{N} h_j I(i \in R_j) \exp\bigl(\mathbf{x}_i^T \boldsymbol{\beta}\bigr) \right\} - n \sum_{j=1}^{d} p_\lambda(|\beta_j|). \end{aligned} \tag{3.5}$$

Taking the derivative with respect to h_j and setting it to be zero, we obtain that

$$\hat{h}_j = \left\{ \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\}^{-1}. \tag{3.6}$$

Substituting $\hat{h}_j$ into (3.5), we get the penalized partial likelihood

$$\sum_{j=1}^{N} \left[\mathbf{x}_{(j)}^T \boldsymbol{\beta} - \log \left\{ \sum_{i \in R_j} \exp\bigl(\mathbf{x}_i^T \boldsymbol{\beta}\bigr) \right\} \right] - n \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{3.7}$$

after dropping a constant term "$-N$". When $p_\lambda(\cdot) \equiv 0$, (3.7) is the partial likelihood function [Cox (1975)]. The penalized likelihood estimate of $\boldsymbol{\beta}$ is derived by maximizing (3.7) with respect to $\boldsymbol{\beta}$. With a proper choice of p_λ, many of the estimated coefficients will be zero and hence their corresponding variables do not appear in the model. This achieves the objectives of variable selection.

3.2. *Frailty model.* It is assumed for the Cox proportional hazards model that the survival times of subjects are independent. This assumption might be violated in some situations, in which the collected data are correlated. One popular approach to model correlated survival times is to use a frailty model. A frailty corresponds to a random block effect that acts multiplicatively on the hazard rates of all subjects in a group. In this section, we only consider the Cox proportional hazard frailty model, in which it is assumed that the hazard rate for the jth subject in the ith subgroup is

$$h_{ij}(t|\mathbf{x}_{ij}, u_i) = h_0(t) u_i \exp\bigl(\mathbf{x}_{ij}^T \boldsymbol{\beta}\bigr), \qquad i = 1, \ldots, n,\ j = 1, \ldots, J_i, \tag{3.8}$$

where the u_i's are associated with frailties, and they are a random sample from some population. It is frequently assumed that given the frailty u_i, the data in the ith group are independent. The most frequently used distribution for frailty is the gamma distribution due to its simplicity. Assume without loss of generality that the mean of frailty is 1 so that all parameters involved are estimable. For the gamma frailty model, the density of u is

$$g(u) = \frac{\alpha^\alpha u^{\alpha-1}\exp(-\alpha u)}{\Gamma(\alpha)}.$$

From (3.1), the full likelihood of "pseudo-data" $\{(u_i, \mathbf{X}_{ij}, Z_{ij}, \delta_{ij}) : i = 1, \ldots, n, j = 1 \ldots, J_i\}$ is

$$\prod_{i=1}^{n}\prod_{j=1}^{J_i}\left[\{h(z_{ij}|\mathbf{x}_{ij}, u_i)\}^{\delta_{ij}}\overline{F}(z_{ij}|\mathbf{x}_{ij}, u_i)\right]\prod_{i=1}^{n} g(u_i).$$

Integrating the full likelihood function with respect to $u_1, \ldots, u_n$, the likelihood of the observed data is given by

$$\begin{aligned} L(\boldsymbol{\beta}, \alpha, H) &= \exp\left\{\boldsymbol{\beta}^T\left(\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}\right)\right\} \\ &\quad\times\prod_{i=1}^{n}\frac{\alpha^\alpha\prod_{j=1}^{J_i}\{h_0(z_{ij})\}^{\delta_{ij}}}{\Gamma(\alpha)\{\sum_{j=1}^{J_i}H_0(z_{ij})\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})+\alpha\}^{(A_i+\alpha)}}, \end{aligned} \tag{3.9}$$

where $A_i = \sum_{j=1}^{J_i}\delta_{ij}$. Therefore the logarithm of the penalized likelihood of the observed data is

$$\begin{aligned} &\sum_{i=1}^{n}\left\{\sum_{j=1}^{J_i}\delta_{ij}\log h(z_{ij}) - \left[(A_i+\alpha)\log\left\{\sum_{j=1}^{J_i}H_0(z_{ij})\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})+\alpha\right\}\right]\right\} \\ &\quad+\sum_{i=1}^{n}\left\{\boldsymbol{\beta}^T\left(\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}\right)+\alpha\log\alpha-\log\Gamma(\alpha)\right\} - n\sum_{j=1}^{d}p_\lambda(|\beta_j|). \end{aligned} \tag{3.10}$$

To eliminate the nuisance parameter $h(\cdot)$, we again employ the profile likelihood method. Consider the "least informative" nonparametric modeling for $H_0(\cdot)$:

$$H_0(z) = \sum_{l=1}^{N}\lambda_l I(z_l \le z), \tag{3.11}$$

where $\{z_1, \ldots, z_N\}$ are pooled observed failure times.

Substituting (3.11) into (3.10), then differentiating it with respect to λ_l, $l = 1, \ldots, N$, the root of the corresponding score function should satisfy the following

equations:

$$\lambda_l^{-1} = \sum_{i=1}^{n} \frac{(A_i + \alpha) \sum_{j=1}^{J_i} I(z_l \le z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})}{\sum_{k=1}^{N} \lambda_k \sum_{j=1}^{J_i} I(z_k \le z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) + \alpha} \tag{3.12}$$

for $l = 1, \ldots, N$.

The above solution does not admit a closed form, and neither does the profile likelihood function. However, the maximum profile likelihood can be implemented as follows. With initial values for α, $\boldsymbol{\beta}$ and λ_l, update $\{\lambda_l\}$ from (3.12) and obtain the penalized profile likelihood of (3.10). With known $H_0(\cdot)$ defined by (3.11), maximize the penalized likelihood (3.10) with respect to $(\alpha, \boldsymbol{\beta})$, and iterate between these two steps. When the Newton–Raphson algorithm is applied to the penalized likelihood (3.10), it involves the first two order derivatives of the gamma function, which may not exist for certain value of α. One approach to avoid this difficulty is the use of a grid of possible values for the frailty parameter α and finding the maxima over this discrete grid, as suggested by Nielsen et al. (1992). Our simulation experience shows that the estimate of $\boldsymbol{\beta}$ is quite empirically robust to the chosen grid of possible values for α. This profile likelihood method appears new even without the task of variable selection. This provides a viable alternative approach to the EM algorithm frequently used in the frailty model.

A natural initial estimator for $\boldsymbol{\beta}$ is the maximum pseudo-partial likelihood estimates of $\boldsymbol{\beta}$ ignoring possible dependency within each group. The corresponding $h_1, \ldots, h_N$ in (3.6) may serve as an initial estimator for $\lambda_1, \ldots, \lambda_N$. Hence given a value of α and initial values of $\boldsymbol{\beta}$ and $\lambda_1, \ldots, \lambda_N$, update the values of $\lambda_1, \ldots, \lambda_N$ and α, $\boldsymbol{\beta}$ in turn until they converge or the penalized profile likelihood of (3.10) fails to change substantially. The proposed algorithm avoids optimizing a high-dimensional problem. It will give us an efficient estimate for $\boldsymbol{\beta}$. The algorithm may converge slowly or even not converge. In this situation, the idea of one-step estimator [see Bickel (1975)] provides us an alternative approach. See Section 3.4 for some other variations.

3.3. *Oracle properties.* We will use the theory of counting processes to establish the oracle property of the proposed variable selection approach for the Cox model under general settings. Following the notation in Andersen and Gill (1982), define $N_i(t) = I\{T_i \le t, T_i \le C_i\}$ and $Y_i(t) = I\{T_i \ge t, C_i \ge t\}$. In this section, the covariate $\mathbf{x}$ is allowed to be time-dependent, denoted by $\mathbf{x}(t)$. For simplicity, we shall work on the finite time interval $[0, \tau]$. Assume without loss of generality that $\tau = 1$. One may extend the results to the interval $[0, \infty)$, following the proof for Theorem 4.2 of Anderson and Gill (1982). We need the following conditions to establish the oracle property.

CONDITIONS. A. $\int_0^1 h_0(t)\, dt < \infty$.

B. The processes $\mathbf{x}(t)$ and $Y(t)$ are left-continuous with right hand limits, and

$$P\{Y(t) = 1 \qquad \forall t \in [0, 1]\} > 0.$$

C. There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ such that

$$E \sup_{t\in[0,1],\boldsymbol{\beta}\in\mathcal{B}} Y(t)\mathbf{x}(t)^T\mathbf{x}(t)\exp\{\boldsymbol{\beta}^T\mathbf{x}(t)\} < \infty.$$

D. Define

$$s^{(0)}(\boldsymbol{\beta},t) = EY(t)\exp\{\boldsymbol{\beta}^T\mathbf{x}(t)\},$$
$$s^{(1)}(\boldsymbol{\beta},t) = EY(t)\mathbf{x}(t)\exp\{\boldsymbol{\beta}^T\mathbf{x}(t)\},$$
$$s^{(2)}(\boldsymbol{\beta},t) = EY(t)\mathbf{x}(t)\mathbf{x}(t)^T\exp\{\boldsymbol{\beta}^T\mathbf{x}(t)\},$$

where $s^{(0)}(\cdot,t)$, $s^{(1)}(\cdot,t)$ and $s^{(2)}(\cdot,t)$ are continuous in $\boldsymbol{\beta}\in\mathcal{B}$, uniformly in $t\in[0,1]$. $s^{(0)}$, $s^{(1)}$ and $s^{(2)}$ are bounded on $\mathcal{B}\times[0,1]$; $s^{(0)}$ is bounded away from zero on $\mathcal{B}\times[0,1]$. The matrix

$$I(\boldsymbol{\beta}_0) = \int_0^1 v(\boldsymbol{\beta}_0,t)s^{(0)}(\boldsymbol{\beta}_0,t)h_0(t)\,dt$$

is finite positive definite, where

$$v(\boldsymbol{\beta},t) = \frac{s^{(2)}}{s^{(0)}} - \left(\frac{s^{(1)}}{s^{(0)}}\right)\left(\frac{s^{(1)}}{s^{(0)}}\right)^T.$$

Conditions A–D guarantee the local asymptotic quadratic (LAQ) property for the partial likelihood function, and hence the asymptotic normality of the maximum partial likelihood estimates. See Andersen and Gill (1982) and Murphy and van der Vaart (2000) for details.

In this section we will show that the proposed estimators perform as well as an oracle estimator. Let $\boldsymbol{\beta}_0 = (\beta_{10},\ldots,\beta_{d0})^T = (\boldsymbol{\beta}_{10}^T,\boldsymbol{\beta}_{20}^T)^T$. Without loss of generality, assume that $\boldsymbol{\beta}_{20} = \mathbf{0}$. Denote by s the number of the components of $\boldsymbol{\beta}_1$,

$$\begin{aligned} a_n &= \max\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0}\neq 0\} \quad \text{and} \\ b_n &= \max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0}\neq 0\}. \end{aligned} \tag{3.13}$$

It will be shown that there exists a penalized partial likelihood estimator that converges at rate $O_P(n^{-1/2}+a_n)$. Oracle properties for the penalized partial likelihood estimator will be also established. In this section, we only state theoretic results. Their proofs will be given in Section 5.

The following theorem shows how the rates of convergence for the penalized partial likelihood estimators depend on the regularization parameter. Let $\ell(\boldsymbol{\beta}) = \sum_{j=1}^N[\mathbf{x}_{(j)}^T\boldsymbol{\beta} - \log\{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}]$ denote the log-partial likelihood function and let $Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^d p_\lambda(|\beta_j|)$ be the penalized partial likelihood function.

THEOREM 3.1. *Assume that* $(\mathbf{x}_1, T_1, C_1), \ldots, (\mathbf{x}_n, T_n, C_n)$ *are independent and identically distributed according to the population* $(\mathbf{x}, T, C)$, T *and* C *are conditionally independent given* $\mathbf{x}$, *and Conditions* (A)–(D) *hold. If* $b_n \to 0$, *then there exists a local maximizer* $\hat{\boldsymbol{\beta}}$ *of* $Q(\boldsymbol{\beta})$ *such that* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$, *where* a_n *is given by* (3.13).

It is clear from Theorem 3.1 that by choosing a proper λ_n, there exists a root-n consistent penalized partial likelihood estimator, as long as $a_n = O(n^{-1/2})$. Denote by

$$\Sigma = \operatorname{diag}\{p''_{\lambda_n}(|\beta_{10}|), \ldots, p''_{\lambda_n}(|\beta_{s0}|)\} \tag{3.14}$$

and

$$\mathbf{b} = \big(p'_{\lambda_n}(|\beta_{10}|)\operatorname{sgn}(\beta_{10}), \ldots, p'_{\lambda_n}(|\beta_{s0}|)\operatorname{sgn}(\beta_{s0})\big)^T, \tag{3.15}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$.

THEOREM 3.2 (Oracle property). *Assume that the penalty function* $p_{\lambda_n}(|\theta|)$ *satisfies condition* (5.6). *If* $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ *and* $a_n = O(n^{-1/2})$, *then under the conditions of Theorem* 3.1, *with probability tending to* 1, *the root-n consistent local maximizer* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ *in Theorem* 3.1 *must satisfy*:

(i) (*Sparsity*) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

(ii) (*Asymptotic normality*)

$$\sqrt{n}(I_1(\boldsymbol{\beta}_{10}) + \Sigma)\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}\mathbf{b}\right\} \to N\{\mathbf{0}, I_1(\boldsymbol{\beta}_{10})\},$$

where $I_1(\boldsymbol{\beta}_{10})$ *is the first* $s \times s$ *submatrix of* $I(\boldsymbol{\beta}_0)$.

Note that for HARD and SCAD, if $\lambda_n \to 0$, then $a_n = 0$ for sufficiently large n. Thus, $\Sigma = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$. Hence when $\sqrt{n}\lambda_n \to \infty$, we have $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\} \to N\{\mathbf{0}, I_1^{-1}(\boldsymbol{\beta}_{10})\}.$$

Therefore HARD and SCAD possess the oracle property when $\lambda_n \to 0$ $\sqrt{n}\lambda_n \to \infty$, and perform as well as the maximum partial likelihood estimates for estimating $\boldsymbol{\beta}_1$ knowing $\boldsymbol{\beta}_2 = \mathbf{0}$. They are more efficient than the maximum partial likelihood estimator for estimating $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

For the L_1-penalty, however, $a_n = \lambda_n$. Hence, the root-n consistency condition in Theorem 3.1 requires that $\lambda_n = O_P(n^{-1/2})$. On the other hand, the oracle property in Theorem 3.2 requires that $\sqrt{n}\lambda_n \to \infty$. Hence, the oracle property does not hold for the LASSO.

Asymptotic properties of the estimators for the regression coefficients in the gamma frailty model have been studied in Parner (1998) and Murphy and van der Vaart (1999) and references therein. Murphy and van der Vaart (2000) established

the LAQ property for the profile likelihood under a general setting. They also illustrated their results for the gamma frailty model when the number J_i of subjects in each group are the same. In what follows, it is assumed that all J_i are the same, denoted by J.

Denote $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T)^T$, and $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0^T)^T$, the true value of $\boldsymbol{\theta}$. Let $PL(\boldsymbol{\theta})$ be the profile likelihood of $L(\boldsymbol{\theta}, H)$ in (3.9). That is,

$$PL(\boldsymbol{\theta}) = \sup_{H \in \mathcal{H}} L(\boldsymbol{\theta}, H),$$

where $\mathcal{H} = \{H : H(z) = \sum_{i=1}^{N} \lambda_l I(z_l \le z)\}$.

Under some regularity conditions, Murphy and van der Vaart (2000) showed that for any random sequence $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$ in probability,

$$\begin{aligned} \log PL(\boldsymbol{\theta}_n) &= \log PL(\boldsymbol{\theta}_0) \\ &\quad + (\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)^T \sum_{i=1}^{n} \tilde{\ell}_0\big\{(\mathbf{x}_{i1}, z_{i1}, \delta_{i1}), \ldots, (\mathbf{x}_{iJ}, z_{iJ}, \delta_{iJ})\big\} \\ &\quad - \tfrac{1}{2} n (\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)^T \tilde{I}_0(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) + o_P\big(\sqrt{n}||\boldsymbol{\theta}_n - \boldsymbol{\theta}_0|| + 1\big)^2, \end{aligned} \tag{3.16}$$

where $\tilde{\ell}_0$ is the *efficient score function* of the marginal likelihood of $\{\mathbf{x}_{i1}, z_{i1}, \delta_{i1}\}$, $\ldots, \{\mathbf{x}_{iJ}, z_{iJ}, \delta_{iJ}\}$ for $\boldsymbol{\theta}$ and $\tilde{I}_0$ the *efficient Fisher information matrix*. To guarantee the existence of a sequence of consistent estimators in (3.17) for the gamma frailty model, one needs to impose some regularity conditions and some bounds on the variance parameters. Those conditions can be found in Parner (1998), in which consistency and asymptotic normality of the nonparametric maximum likelihood estimator were investigated.

When the LAQ property (3.17) holds, we may establish the oracle property for the penalized profile likelihood for the gamma frailty model. Here we only state the results. Their proofs are given in Section 5. Denote the logarithm of penalized profile likelihood by $Q(\boldsymbol{\theta}) = \log PL(\boldsymbol{\theta}) - n \sum_{j=1}^{d} p_\lambda(|\beta_j|)$.

THEOREM 3.3. *Assume that* $(\mathbf{x}_{ij}, T_{ij}, C_{ij})_{j=1}^{J}$ *are independent random samples for* $i = 1, \ldots, n$, *and given* u_i, $(\mathbf{x}_{ij}, T_{ij}, C_{ij})$, $j = 1, \ldots, J$, *are independently distributed according to* (3.8). T_{ij} *and* C_{ij} *are conditionally independent given* $\mathbf{x}_i$ *and* $\{u_i\}$ *are i.i.d. from a Gamma distribution. If* $b_n \to 0$ *and the local asymptotic quadratic property* (3.17) *holds, then there exists a local maximizer* $\hat{\boldsymbol{\theta}}$ *of* $Q(\boldsymbol{\theta})$ *such that* $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_P(n^{-1/2} + a_n)$, *where* a_n *is given by* (3.13).

To state the oracle properties, let

$$\hat{\boldsymbol{\theta}}_1 = \big(\hat{\alpha}, \hat{\boldsymbol{\beta}}_1^T\big)^T, \ \boldsymbol{\theta}_{10} = \big(\alpha_0, \boldsymbol{\beta}_{10}^T\big)^T, \ \ \hat{\boldsymbol{\theta}} = \big(\hat{\boldsymbol{\theta}}_1^T, \hat{\boldsymbol{\beta}}_2^T\big)^T \quad \text{and} \quad \boldsymbol{\theta}_0 = \big(\boldsymbol{\theta}_{10}^T, \boldsymbol{\beta}_{20}^T\big)^T.$$

Denote by

$$\Sigma_1 = \operatorname{diag}(0, \Sigma) \quad \text{and} \quad \mathbf{b}_1 = \bigl(0, \mathbf{b}^T\bigr)^T$$

with Σ and $\mathbf{b}$ given in (3.14) and (3.15). Now we state the oracle properties of $\hat{\boldsymbol{\theta}}$.

THEOREM 3.4. *Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies condition* (5.6). *If $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ and $a_n = O(n^{-1/2})$, then under the conditions of Theorem* 3.3, *with probability tending to* 1, *the root-n consistent local maximizer $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ in Theorem* 3.3 *must satisfy*:

(i) (*Sparsity*) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

(ii) (*Asymptotic normality*)

$$\sqrt{n}\bigl(\tilde{I}_1(\boldsymbol{\theta}_{10}) + \Sigma_1\bigr)\bigl\{\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10} + \bigl(\tilde{I}_1(\boldsymbol{\theta}_{10}) + \Sigma_1\bigr)^{-1}\mathbf{b}_1\bigr\} \to N\bigl\{\mathbf{0}, \tilde{I}_1(\boldsymbol{\theta}_{10})\bigr\},$$

where $\tilde{I}_1(\boldsymbol{\theta}_{10})$ consists of the first $(s+1) \times (s+1)$ submatrix of $\tilde{I}_0(\boldsymbol{\theta}_{10}, \mathbf{0})$.

With a proper choice of regularization parameter λ_n, the penalized likelihood estimators with a class of penalty functions possess the oracle property under some mild regularity conditions. In practice, data-driven methods, such as cross validation and generalized cross validation, are employed to select λ_n. For a linear estimator (in terms of response variable), asymptotic optimal properties of such choice of λ_n have been studied in series of papers by Wahba (1985) and Li (1987) and references therein. With the local quadratic approximations in Section 3.4, the resulting estimators will be approximately locally linear. As pointed out by a referee, it is of interest to establish the asymptotic property of the proposed estimators with a data-driven λ_n. Further studies on this issue are needed, but it is beyond the scope of this paper.

3.4. *Local quadratic approximations and standard errors.* Note that the penalty function $p_\lambda(|\beta_j|)$ is irregular at the origin and may not have a second derivative at some points. Some special care is needed before applying the Newton–Raphson algorithm. Following Fan and Li (2001a), we locally approximate the penalty functions introduced in Section 2 by quadratic functions as follows. Given an initial value $\boldsymbol{\beta}_0$ that is close to the maximizer of the penalized likelihood function, when β_{j0} is not very close to 0, the penalty $p_\lambda(|\beta_j|)$ can be locally approximated by the quadratic function as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\operatorname{sgn}(\beta_j) \approx \bigl\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\bigr\}\beta_j,$$

otherwise, set $\hat{\beta}_j = 0$. In other words,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \tfrac{1}{2}p'_\lambda(|\beta_{j0}|)\bigl(\beta_j^2 - \beta_{j0}^2\bigr) \qquad \text{for } \beta_j \approx \beta_{j0}.$$

Similarly, approximate the profile likelihood via Taylor's expansion. The maximization can be reduced to a local quadratic maximization problem. This results in a modified Newton–Raphson algorithm.

As in the maximum likelihood estimation setting, with a good initial value $\boldsymbol{\beta}_0$, the one-step penalized partial likelihood estimator can be as efficient as the fully iterative one, namely, the penalized maximum likelihood estimate, when one uses the Newton–Raphson algorithm [see Bickel (1975)]. Furthermore estimators obtained after a few iterations can be always regarded as one-step estimators, which is as efficient as the fully iterative method. Indeed, Robinson (1988) shows the rate of convergence for the difference between a finite-step estimator and the fully-iterative MLE. In this sense, one does not have to iterate the algorithm until convergence as long as the initial estimators are good enough.

The standard errors for estimated parameters can be directly obtained because we are estimating parameters and selecting variables at the same time. Following the conventional technique in the likelihood setting, the corresponding sandwich formula can be used as an estimator for the covariance matrix of the estimates $\hat{\boldsymbol{\beta}}$. For the Cox proportional hazards model, the solution in the Newton–Raphson algorithm is updated by

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \{\nabla^2 \ell(\boldsymbol{\beta}_0) - n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\{\nabla \ell(\boldsymbol{\beta}_0) - n\mathbf{U}_\lambda(\boldsymbol{\beta}_0)\}, \tag{3.17}$$

where $\ell(\boldsymbol{\beta})$ is the partial likelihood

$$\nabla \ell(\boldsymbol{\beta}_0) = \frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}, \quad \nabla^2 \ell(\boldsymbol{\beta}_0) = \frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T},$$

$$\Sigma_\lambda(\boldsymbol{\beta}_0) = \operatorname{diag}\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \ldots, p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\} \quad \text{and}$$
$$\mathbf{U}_\lambda(\boldsymbol{\beta}_0) = \Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0.$$

Thus the corresponding sandwich formula is given by

$$\widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}}) = \{\nabla^2 \ell(\hat{\boldsymbol{\beta}}) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1} \widehat{\operatorname{cov}}\{\nabla \ell(\hat{\boldsymbol{\beta}})\}\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}. \tag{3.18}$$

This formula is consistent with Theorem 3.2 and will be shown to have good accuracy for moderate sample sizes. The sandwich formula for the frailty model can be derived in the same way.

4. Simulation studies and applications.

4.1. *Selection of thresholding parameters.* To implement the methods described in previous sections, it is desirable to have an automatic method for selecting the thresholding parameter λ involved in $p_\lambda(\cdot)$ based on data. Here we estimate λ via minimizing an approximate generalized cross-validation (GCV) statistic [Craven and Wahba (1977)]. Regarding the penalized partial likelihood

as an iteratively reweighted least-squares problem, by some straightforward calculation, the effective number of parameters for the Cox proportional hazards model in the last step of the Newton–Raphson algorithm iteration is

$$e(\lambda) = \operatorname{tr}\bigl[\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}) + \Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1} \nabla^2 \ell(\hat{\boldsymbol{\beta}})\bigr].$$

Therefore the generalized cross-validation statistic is defined by

$$\mathrm{GCV}(\lambda) = \frac{-\ell(\hat{\boldsymbol{\beta}})}{n\{1 - e(\lambda)/n\}^2}$$

and $\hat{\lambda} = \operatorname{argmin}_\lambda\{\mathrm{GCV}(\lambda)\}$ is selected. Similarly the corresponding generalized cross-validation statistic can be defined for the penalized profile likelihood function for the frailty model (3.8).

4.2. *Prediction and model error.* When the covariate $\mathbf{x}$ is random, if $\hat{\mu}(\mathbf{x})$ is a prediction procedure constructed using the present data, the prediction error is defined as

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

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

$$\mathrm{PE}(\hat{\mu}) = E\,\mathrm{Var}(Y|\mathbf{x}) + E\bigl\{E(Y|\mathbf{x}) - \hat{\mu}(\mathbf{x})\bigr\}^2.$$

The first component is inherently 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 $\mathrm{ME}(\hat{\mu})$. For the Cox proportional hazards model (3.2),

$$\mu(\mathbf{x}) = E(T|\mathbf{x}) = \int_0^\infty t h_0(t) \exp\bigl(\mathbf{x}^T\boldsymbol{\beta}\bigr) \exp\biggl\{-\int_0^t h_0(u) \exp\bigl(\mathbf{x}^T\boldsymbol{\beta}\bigr)\,du\biggr\}\,dt.$$

In the following simulation examples, it will be taken that $h_0(t) \equiv 1$. Thus by some algebra calculation,

$$\mu(\mathbf{x}) = \exp\bigl(-\mathbf{x}^T\boldsymbol{\beta}\bigr).$$

For the Cox frailty model with $h_0(t) \equiv 1$,

$$\mu(\mathbf{x}) = \exp\bigl(-\mathbf{x}^T\boldsymbol{\beta}\bigr) E\bigl(u^{-1}\bigr).$$

The factor $E(u^{-1})$, due to the frailty, is dropped off when the performance of two different approaches is compared in terms of their Relative Model Errors (RME), defined as the ratio of the model errors of the two approaches. Therefore, the model error will be defined as

$$E\bigl\{\exp\bigl(-\mathbf{X}^T\hat{\boldsymbol{\beta}}\bigr) - \exp\bigl(-\mathbf{X}^T\boldsymbol{\beta}_0\bigr)\bigr\}^2$$

for both the Cox model and the frailty model.

4.3. *Simulations.* In the following examples, we numerically compare the proposed variable selection methods with the maximum partial likelihood estimate and the best subset variable selection. All simulations are conducted using MATLAB codes. To find the best subset variable selection, we searched exhaustively over all possible subsets and selected the subset with the best BIC score.

EXAMPLE 1. In this example we simulated 100 data sets consisting of $n = 75$ and 100 observations from the exponential hazard model

$$h(t|\mathbf{x}) = \exp(\mathbf{x}^T \boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^T$. The x_i were marginally standard normal and the correlation between x_i and x_j was $\rho^{|i-j|}$ with $\rho = 0.5$. The distribution of the censoring time is an exponential distribution with mean $U \exp(\mathbf{x}^T \boldsymbol{\beta}_0)$, where U is randomly generated from the uniform distribution over $[1, 3]$ for each simulated data set so that about 30% data are censored. Here $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$ which is regarded as a known constant so that the censoring scheme is noninformative. This model will give us that the standard error of the maximum partial likelihood estimator $\hat{\beta}_7$ is about 0.3 when $n = 75$, resulting in a t-statistic around 2. This is a challenge to any data-driven variable selection technique on whether or not to include this variable in the model.

Model errors of the proposed procedures are compared to those of the maximum partial likelihood estimates. Following Tibshirani (1996), we compare the Median of Relative Model Errors (MRME) rather than the mean of relative model errors due to instability of the best subset variable selection and HARD. The MRME over the 100 simulated data sets summarized in Table 1. The average number of zero coefficients is also reported in Table 1, in which the column labeled "correct"

TABLE 1
Simulation results for Cox's proportional hazards model

Method	MRME(%)	Aver. no. of 0 coeff.	
		correct	incorrect
$n = 75$			
SCAD	0.3696	4.34	0.18
LASSO	0.4559	4.05	0.10
HARD	0.3866	4.87	0.29
Best subset	0.4505	4.75	0.15
Oracle	0.3115	5	0
$n = 100$			
SCAD	0.3346	4.33	0.03
LASSO	0.4582	3.99	0.02
HARD	0.4367	4.84	0.08
Best subset	0.4624	4.76	0.06
Oracle	0.3369	5	0

TABLE 2
Standard deviations for the Cox proportional hazards models ($n = 100$)

Method	$\hat{\beta}_1$		$\hat{\beta}_4$		$\hat{\beta}_7$	
	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD	0.191	0.154(0.023)	0.143	0.169(0.020)	0.188	0.141(0.026)
LASSO	0.190	0.150(0.019)	0.163	0.143(0.016)	0.157	0.118(0.016)
HARD	0.180	0.161(0.019)	0.155	0.175(0.014)	0.144	0.147(0.015)
Best subset	0.184	0.161(0.019)	0.154	0.175(0.015)	0.141	0.148(0.015)
Oracle	0.173	0.156(0.019)	0.165	0.169(0.017)	0.137	0.147(0.015)

presents the average restricted only to the true zero coefficients, while the column labeled "incorrect" depicts the average of coefficients erroneously set to 0. From Table 1, the SCAD outperforms the other three methods and performs as well as the oracle estimator in terms of MRME. All methods select about the same correct number of significant variables.

We now test the accuracy of the proposed standard error formula. The median absolute deviation divided by 0.6745, denoted by SD in Table 2, of the 100 estimated coefficients in the 100 simulations can be regarded as the true standard error except the Monte Carlo error. The median of the 100 estimated SDs, resulting from 100 simulations, denoted by SD_m, and the median absolute deviation error of the 100 estimated standard errors divided by 0.6745, denoted by SD_{mad}, 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 excluded from the selected model. Table 2 only presents the results for non-zero coefficients when the sample size $n = 100$. The results for the other two cases with $n = 75$ are similar. The last row of Table 2 lists the standard deviations and standard errors for the oracle estimator, obtained by fitting the ideal model consisting of variables X_1, X_4 and X_7. Table 2 suggests that the proposed standard error formula performs well, and SCAD and HARD perform as well as the oracle estimator in terms of estimating the standard errors of the estimated coefficients.

EXAMPLE 2. In this example we simulated 100 data sets consisting of n groups and J subjects in each group from the exponential hazard frailty model

$$h(t|\mathbf{x}, u) = u \exp(\mathbf{x}^T \boldsymbol{\beta}),$$

where the $\boldsymbol{\beta}$ and $\mathbf{x}$ are the same as those in Example 1, and the frailty u is the gamma frailty with $\alpha = 4$.

The performance of variable selection via nonconcave penalized likelihood and the best subset variable selection is compared in terms of their model errors, model complexity and accuracy. Model errors of the proposed procedures are compared to those of the maximum profile likelihood estimates. The Median of Relative

TABLE 3
Simulation results for frailty model

Method	MRME(%)	Aver. no. of 0 coeff.	
		correct	incorrect
$n = 50,\ J = 2$			
SCAD	0.5322	4.18	0.14
LASSO	0.8880	4.04	0.06
Hard	0.5784	4.54	0.09
Best Subset	0.4251	4.78	0.07
Oracle	0.3592	5	0
$n = 75,\ J = 2$			
SCAD	0.5177	4.18	0
LASSO	1.4075	4.08	0
Hard	0.5782	4.50	0
Best Subset	0.5188	4.89	0
Oracle	0.4886	5	0
$n = 100,\ J = 2$			
SCAD	0.4930	4.29	0
LASSO	1.0438	4.10	0
Hard	0.6379	4.42	0
Best subset	0.6019	4.85	0
Oracle	0.5631	5	0

Model Errors (MRME) over 100 simulated data sets with some combinations of n and J is summarized in Table 3, and the standard errors for estimated nonzero coefficients with $n = 100$ and $J = 2$ are depicted in Table 4. The last row of Table 4 displays the standard deviations and standard errors of estimated coefficients based on the true model (oracle estimate). From Tables 3 and 4, SCAD performs as well as the oracle estimator and outperforms the other approaches. Comparing the standard deviations for the SCAD and those for the oracle estimate in Table 4, it can be seen that the SCAD performs as well as if one knew the true model in advance.

TABLE 4
Standard deviations for frailty models ($n = 100,\ J = 2$)

	$\hat{\beta}_1$		$\hat{\beta}_4$		$\hat{\beta}_7$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD	0.114	0.100(0.012)	0.092	0.095(0.007)	0.113	0.098(0.008)
LASSO	0.098	0.077(0.007)	0.086	0.082(0.006)	0.097	0.072(0.008)
Hard	0.083	0.101(0.008)	0.095	0.102(0.008)	0.094	0.102(0.009)
Best Subset	0.080	0.103(0.008)	0.092	0.103(0.008)	0.090	0.102(0.010)
Oracle	0.083	0.100(0.008)	0.089	0.102(0.007)	0.087	0.102(0.009)

EXAMPLE 3. The proposed approach is now applied to the "nursing home" data set analyzed by Morris, Norton and Zhou (1994), where a full description of this data set is given. Here is a brief summary. The data were from an experiment sponsored by the National Center for Health Services Research in 1980–1982, involving 36 for-profit nursing homes in San Diego, California. The experiment was designed to assess the effects of differing financial incentives on the admission of nursing home patients, on their subsequent care, and on the durations of stay. The 18 treatment nursing homes received higher per diem payments for accepting more disabled medicaid patients. They also received bonuses for improving a patient's health status and for discharging patients to their homes within 90 days. These incentives were not offered to the 18 control nursing homes. Altogether 1601 samples are available.

Morris, Norton and Zhou (1994) took *days in the nursing home* as the response variable t. They suggested the use of the following model:

$$h(t|\mathbf{x}) = h_0(t)\exp\left(\sum_{i=0}^{7} x_i\beta_i\right), \tag{4.1}$$

where x_1 is a treatment indicator, being 1 if treated at a nursing home and 0 otherwise; x_2 is the variable age, which ranges from 65 to 90; x_3 is a gender variable, being 1 if male and 0 if female; x_4 is a marital status indicator, being 1 if married and 0 otherwise; x_5 , x_6 and x_7 are three binary health status indicators, corresponding from the best health to the worst health. The parameter β_0 is an intercept when a parametric model for the baseline h_0 is employed, while it is dropped from the model if the nonparametric model for h_0 is used. Morris, Norton and Zhou (1994) fitted the Cox model with three parametric and the nonparametric baseline hazard models to this data set. Their model does not include any possible interactions. To explore possible interaction and to reduce possible modeling biases, all interactions among treatment, age, gender and marital status are included in the initial model, and fit to the data by the Cox regression model with 13 covariates. Only x_2 is standardized as other variables are binary. Penalized partial likelihood approach with the SCAD, L_1 and hard penalty are applied to this data set. The thresholding parameter λ, selected by the GCV, is 0.0227, 0.0113 and 0.0890 for the SCAD, LASSO and HARD, respectively. The best subset variable selection with AIC and BIC is also conducted. Estimated coefficients and their standard errors are shown in Table 5.

From Table 5, the best subset variable selection with AIC and the SCAD yield almost the same model. Compared with the other approaches, the LASSO somewhat shrinks all nonzero coefficients, while the best subset variable selection with BIC results in too simple a model as it over-penalizes the dimension of the selected model.

The resulting models are somewhat different from the one without including interactions, presented by Morris, Norton and Zhou (1994). The main differences are summarized as follows.

TABLE 5
Estimated coefficients and standard errors

	MLE	Best (BIC)	Best (AIC)	SCAD	LASSO	HARD
TRT	-0.04(0.07)	0(−)	0(−)	0(−)	0(−)	0(−)
Age	-0.12(0.05)	0(−)	-0.09(0.03)	-0.09(0.04)	-0.05(0.02)	0(−)
Gender	0.43(0.10)	0.40(0.06)	0.44(0.08)	0.44(0.08)	0.31(0.05)	0.44(0.08)
Married	0.22(0.14)	0(−)	0.16(0.08)	0.18(0.08)	0.08(0.03)	0.18(0.08)
Health1	0.03(0.08)	0(−)	0(−)	0(−)	0(−)	0(−)
Health2	0.24(0.07)	0.24(0.06)	0.23(0.06)	0.23(0.06)	0.14(0.04)	0.23(0.06)
Health3	0.57(0.10)	0.53(0.09)	0.54(0.09)	0.54(0.09)	0.35(0.06)	0.55(0.09)
TRT*Age	0.07(0.06)	0(−)	0(−)	0(−)	0(−)	0(−)
TRT*Gender	-0.10(0.13)	0(−)	-0.15(0.11)	-0.16(0.11)	0(−)	-0.15(0.11)
TRT*Married	-0.00(0.16)	0(−)	0(−)	0(−)	0(−)	0(−)
Age*Gender	0.16(0.06)	0(−)	0.17(0.06)	0.16(0.06)	0.07(0.03)	0.09(0.05)
Age*Married	0.09(0.08)	0(−)	0(−)	0.09(0.08)	0(−)	0(−)
Gender*Married	-0.07(0.16)	0(−)	0(−)	0(−)	0(−)	0(−)

In the model excluding interactions, the *age* variable is not statistically significant, pointed out by Morris, Norton and Zhou (1994). However, it is very significant in the resulting model with interactions. It is clear from Table 5 that elderly patients are more likely (less risky) stay at nursing home.

From Table 5, the interaction between the variables *treatment* and *gender* is selected by SCAD and HARD, although treatment is not significant. It seems that men prefer to stay at a nursing home with treatment, while elderly men like to leave a nursing home earlier. The latter is because elderly men are much more likely to be married [see Morris, Norton and Zhou (1994)], and they like to stay at their own home rather than a nursing home.

5. Proofs. In this section, we give rigorous proofs of Theorems 3.1 and 3.2 and sketch proof of Theorems 3.3 and 3.4. Rigorous proofs of Theorems 3.3 and 3.4 can be found in an earlier version of this paper [Fan and Li (2001b)].

PROOF OF THEOREM 3.1. The partial likelihood $\ell(\boldsymbol{\beta})$ can be written as

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \int_0^1 \boldsymbol{\beta}^T \mathbf{X}_i(s)\, dN_i(s) \\ &\quad - \int_0^1 \log\left\{ \sum_{i=1}^{n} Y_i(s) \exp\bigl(\boldsymbol{\beta}^T \mathbf{X}_i(s)\bigr) \right\} d\bar{N}(s),\end{aligned} \tag{5.1}$$

where $\bar{N} = \sum_{i=1}^{n} N_i$. Using Theorem 4.1 and the proof of Lemma 3.1 of Andersen and Gill (1982) [see also Theorem VII 2.1 of Anderson, Borgan, Gill and Keiding

(1993)], it follows that under Conditions A–D for each $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$,

$$\begin{aligned}\frac{1}{n}\{\ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}_0)\} = \int_0^1 \Bigg[&(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s^{(1)}(\boldsymbol{\beta}_0, t) \\ &- \log\left\{\frac{s^{(0)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}_0, t)}\right\} s^{(0)}(\boldsymbol{\beta}_0, t)\Bigg] h_0(t)\, dt \\ &+ O_P\left(\frac{||\boldsymbol{\beta} - \boldsymbol{\beta}_0||}{\sqrt{n}}\right).\end{aligned} \tag{5.2}$$

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

$$P\left\{\sup_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) < Q(\boldsymbol{\beta}_0)\right\} \geq 1 - \varepsilon. \tag{5.3}$$

This implies 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 $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\alpha_n)$.

Using $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\cdot) \geq 0$, we have

$$\begin{aligned}D_n(\mathbf{u}) &\equiv \frac{1}{n}\{Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - Q(\boldsymbol{\beta}_0)\} \\ &\leq \frac{1}{n}\{\ell(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\beta}_0)\} \\ &\quad - \sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\},\end{aligned} \tag{5.4}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$. By (5.3) and Taylor's expansion, we have

$$\begin{aligned}&\frac{1}{n}\{\ell(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\beta}_0)\} \\ &\quad = -\frac{1}{2}\alpha_n^2 \mathbf{u}^T\{I(\boldsymbol{\beta}_0) + o_P(1)\}\mathbf{u} + O_P\big(n^{-1/2}\alpha_n \|\mathbf{u}\|\big),\end{aligned} \tag{5.5}$$

as the first order derivative of the first term in (5.3) equals 0. Since $I(\boldsymbol{\beta}_0)$ is positive definite, the first term in the right-hand side of (5.6) is of the order $C^2\alpha_n^2$. Note that $n^{-1/2}\alpha_n = O_p(\alpha_n^2)$. By choosing a sufficiently large C, the first term in the last equation will dominate the second term, uniformly in $||\mathbf{u}|| = C$. On the other hand, by the Taylor expansion and the Cauchy–Schwarz inequality, the second term in the right-hand side of (5.4) is bounded by

$$\sqrt{s}\alpha_n a_n ||\mathbf{u}|| + \alpha_n^2 b_n ||\mathbf{u}||^2 = C\alpha_n^2\big(\sqrt{s} + b_n C\big)$$

which is dominated by the first term of (5.6) as $b_n \to 0$, when C is sufficiently large. Hence by choosing sufficiently large C, (5.3) holds. This completes the proof of the theorem. $\square$

To establish the oracle property, we show that this estimator must process the sparsity property $\hat{\boldsymbol{\beta}}_2 = 0$, which is stated as follows.

LEMMA 5.1. *Assume that*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0, \tag{5.6}$$

$\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$, *and the conditions of Theorem* 3.1 *hold. Then 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,

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

PROOF. It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying that $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_P(n^{-1/2})$, and $\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}$, $\partial Q(\boldsymbol{\beta})/\partial\beta_j$ and β_j have different signs for $\beta_j \in (-Cn^{-1/2}, Cn^{-1/2})$ for $j = s+1, \ldots, d$. From (5.3), for each $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$, we have

$$\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}_0) + nf(\boldsymbol{\beta}) + O_P\big(\sqrt{n}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|\big),$$

where

$$f(\boldsymbol{\beta}) = \int_0^1 \left[(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s^{(1)}(\boldsymbol{\beta}_0, t) - \log\left\{\frac{s^{(0)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}_0, t)}\right\} s^{(0)}(\boldsymbol{\beta}_0, t)\right] h_0(t)\, dt.$$

Note that

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \int_0^1 \left[\frac{s^{(1)}(\boldsymbol{\beta}_0, t)}{s^{(0)}(\boldsymbol{\beta}_0, t)} - \frac{s^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}\right] s^{(0)}(\boldsymbol{\beta}_0, t) h_0(t)\, dt$$

and

$$-\frac{\partial^2 f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \int_0^1 \left[\frac{s^{(2)}(\boldsymbol{\beta}, t) s^{(0)}(\boldsymbol{\beta}, t) - s^{(1)}(\boldsymbol{\beta}, t)\{s^{(1)}(\boldsymbol{\beta}, t)\}^T}{[s^{(0)}(\boldsymbol{\beta}, t)]^2}\right] s^{(0)}(\boldsymbol{\beta}_0, t) h_0(t)\, dt.$$

Thus

$$f(\boldsymbol{\beta}_0) = 0, \qquad \left.\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = 0$$

and

$$\left. -\frac{\partial^2 f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = I(\boldsymbol{\beta}_0)$$

which is a finite positive definite matrix. Therefore for each $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$

$$f(\boldsymbol{\beta}) = -\tfrac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\{I(\boldsymbol{\beta}_0) + o(1)\}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

By the Taylor expansion, for $\boldsymbol{\beta}$ in a $n^{-1/2}$-neighborhood of $\boldsymbol{\beta}_0$, we have

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} - np'_{\lambda_n}(|\beta_j|)\operatorname{sgn}(\beta_j)\\ &= -n\sum_{l=1}^{d}\frac{\partial^2 f(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_l}(\beta_l - \beta_{l0}) + O_P\bigl(n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2\bigr) - np'_{\lambda_n}(|\beta_j|)\operatorname{sgn}(\beta_j)\\ &= -np'_{\lambda_n}(|\beta_j|)\operatorname{sgn}(\beta_j) + O_P\bigl(n^{1/2}\bigr),\end{aligned}$$

where $I_{jl}(\boldsymbol{\beta}_0)$ is the (j,l)-element of $I(\boldsymbol{\beta}_0)$. Thus, it follows that

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_n\bigl\{-\lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\operatorname{sgn}(\beta_j) + O_P\bigl(n^{-1/2}/\lambda_n\bigr)\bigr\}.$$

Since $\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\lambda_n^{-1}p'_{\lambda_n}(\theta) > 0$ and $n^{-1/2}/\lambda_n \to 0$, the sign of the derivative is completely determined by that of β_j. This completes the proof. □

PROOF OF THEOREM 3.2. It follows by Lemma 5.1 that Part (i) holds. Now we prove Part (ii). Using the proof of Theorem 3.1, it can be shown that there exists a $\hat{\boldsymbol{\beta}}_1$ in Theorem 3.1 that is a root-n consistent local maximizer of $Q\{(\boldsymbol{\beta}_1^T, \mathbf{0})^T\}$, satisfying the likelihood equations

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_1^T,\mathbf{0})^T} = 0 \qquad \text{for } j = 1,\ldots,s. \tag{5.7}$$

Let $\mathbf{U}(\boldsymbol{\beta})$ be the score function of (5.1), that is

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^1 \mathbf{X}_i(s)\,dN_i(s) - \int_0^1 \frac{\sum_{i=1}^n Y_i(s)\mathbf{X}_i(s)\exp\{\boldsymbol{\beta}^T\mathbf{X}_i(s)\}}{\sum_{i=1}^n Y_i(s)\exp\{\boldsymbol{\beta}^T\mathbf{X}_i(s)\}}\,d\bar{N}(s),$$

and denote

$$\begin{aligned}\hat{I}(\boldsymbol{\beta}) = \int_0^1 \Biggl(&\frac{\sum_{i=1}^n Y_i(s)\mathbf{X}_i(s)\mathbf{X}^T(s)\exp\{\boldsymbol{\beta}^T\mathbf{X}_i(s)\}}{\sum_{i=1}^n Y_i(s)\exp\{\boldsymbol{\beta}^T\mathbf{X}_i(s)\}}\\ &- \frac{[\sum_{i=1}^n Y_i\mathbf{X}_i(s)\exp\{\boldsymbol{\beta}^T\mathbf{X}_i(s)\}][\sum_{i=1}^n Y_i\mathbf{X}_i(s)\exp\{\boldsymbol{\beta}^T\mathbf{X}_i(s)\}]^T}{[\sum_{i=1}^n Y_i(s)\exp\{\boldsymbol{\beta}^T\mathbf{X}_i(s)\}]^2}\Biggr)d\bar{N}(s).\end{aligned}$$

Note that $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator and $\beta_{j0} \neq 0$. By Taylor's expansion, it holds for $j = 1,\ldots,s$ that

$$\begin{aligned}
&\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_1^T,\mathbf{0})^T} - np'_{\lambda_n}(|\hat{\beta}_j|) \\
&\quad = U_j(\boldsymbol{\beta}_0) - \sum_{l=1}^{s} \hat{I}_{jl}(\boldsymbol{\beta}^*)(\hat{\beta}_l - \beta_{l0}) \\
&\qquad - n\left(p'_{\lambda_n}(|\beta_{j0}|)\operatorname{sgn}(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + o_P(1)\}(\hat{\beta}_j - \beta_{j0})\right),
\end{aligned}$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$, $U_j(\boldsymbol{\beta}_0)$ is the j-th component of $\mathbf{U}(\boldsymbol{\beta}_0)$ and $\hat{I}_{jl}(\boldsymbol{\beta}^*)$ is the (j,l)- element of $\hat{I}(\boldsymbol{\beta}^*)$.

Using Theorem 3.2 of Andersen and Gill (1982), it can be proved that

$$\frac{1}{\sqrt{n}}\mathbf{U}_1(\boldsymbol{\beta}_0) \to N\{\mathbf{0}, I_1(\boldsymbol{\beta}_0)\}$$

in distribution as $n \to \infty$, where $\mathbf{U}_1(\boldsymbol{\beta}_0)$ consists of the first s elements of $\mathbf{U}(\boldsymbol{\beta}_0)$, and $I_1(\boldsymbol{\beta}_{10})$ consists of the first s rows and columns of $I(\boldsymbol{\beta}_0)$; furthermore,

$$\frac{1}{n}\hat{I}(\boldsymbol{\beta}^*) \to I_1(\boldsymbol{\beta}_0)$$

in probability as $n \to \infty$. Since $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \mathbf{0})^T$, it follows by using Slutsky's Theorem that

$$\sqrt{n}(I_1(\boldsymbol{\beta}_{10}) + \Sigma)\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \left(I_1(\boldsymbol{\beta}_{10}) + \Sigma\right)^{-1}\mathbf{b}\right\} \to N\left\{\mathbf{0}, I_1(\boldsymbol{\beta}_{10})\right\}.$$

This completes the proof. □

PROOF OF THEOREM 3.3. Denote $\alpha_n = n^{-1/2} + a_n$, and let $\alpha_n \to 0$. It follows that for any $\mathbf{u}$ with $\|\mathbf{u}\| = C$, $\boldsymbol{\theta}_0 + \alpha_n\mathbf{u} \to \boldsymbol{\theta}_0$. Therefore (3.17) holds for $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \alpha_n\mathbf{u}$, which implies that

$$\begin{aligned}
&\frac{1}{n}\{\log PL(\boldsymbol{\theta}_0 + \alpha_n\mathbf{u}) - \log PL(\boldsymbol{\theta}_0)\} \\
&\quad = -\frac{1}{2}\alpha_n^2\mathbf{u}^T\tilde{I}_0(\boldsymbol{\theta}_0)\mathbf{u} \\
&\qquad + \frac{\alpha_n\mathbf{u}^T}{n}\sum_{i=1}^{n}\tilde{\ell}_0\{(\mathbf{x}_{i1}, z_{i1}, \delta_{i1}), \ldots, (\mathbf{x}_{iJ}, z_{iJ}, \delta_{iJ})\} \\
&\qquad + o_P\left(\alpha_n\|\mathbf{u}\| + \frac{1}{\sqrt{n}}\right)^2.
\end{aligned} \tag{5.8}$$

As $\tilde{\ell}_0\{(\mathbf{x}_{i1}, z_{i1}, \delta_{i1}), \ldots, (\mathbf{x}_{iJ}, z_{iJ}, \delta_{iJ})\}$ is the efficient score function of marginal likelihood of the i-th group data at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the second term of (5.9) is of the order $O_P(\alpha_n\|\mathbf{u}\|/\sqrt{n})$. Note that $\alpha_n/\sqrt{n} = O_P(\alpha_n^2)$. By choosing sufficient large

C, the first term will dominate the second one, uniformly in $\|\mathbf{u}\| = C$. Following the same strategy as the proof of Theorem 3.1, it can be shown that the results in Theorem 3.3 hold. We omit the details, but see Fan and Li (2001b) for a rigorous proof. □

PROOF OF THEOREM 3.4. The sparsity in Part (i) can be established following the same lines in the proof of Lemma 5.1.

Similarly to the proof of Theorem 3.2, it follows by Corollary 1 of Murphy and van der Vaart (2000) that

$$\sqrt{n}\{\tilde{I}_1(\boldsymbol{\theta}_{10}) + \Sigma_1\}[\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10} + \{\tilde{I}_1(\boldsymbol{\theta}_{10}) + \Sigma_1\}^{-1}\mathbf{b}_1] \to N(\mathbf{0}, \tilde{I}_1(\boldsymbol{\theta}_{10}))$$

in distribution, where $\tilde{I}_1(\boldsymbol{\theta}_{10})$ consists of the first $(s+1) \times (s+1)$ submatrix of $\tilde{I}_0(\boldsymbol{\theta}_{10}, \mathbf{0})$. See Fan and Li (2001b) for a rigorous proof. □

Acknowledgments. The authors would like to thank the referees for constructive comments that led to improvement an earlier draft of the paper.

REFERENCES

ANDERSEN, P. K, BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

ANTONIADIS, A. (1997). Wavelets in Statistics: A review (with discussion). *J. Italian Statist. Assoc.* **6** 97–144.

ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations (with discussion). *J. Amer. Statist. Assoc.* **96** 939–967.

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434.

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

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

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.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

FAN, J. (1997). Comment on "Wavelets in statistics: a review" by A. Antoniadis. *J. Italian Statist. Assoc.* **6** 131–138.

FAN, J. and LI, R. (2001a). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

FAN, J. and LI, R. (2001b). Variable selection for Cox's proportional hazards model and frailty model. Institute of Statistic Mimeo Series #2372, Dept. Statistics, Univ. North Carolina, Chapel Hill.

FARAGGI, D. and SIMON, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54** 1475–1485.

KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

LI, K. C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross validation: discrete index set. *Ann. Statist.* **15** 958–975.

LINDLEY, D. V. (1968). The choice of variables in multiple regression (with discussion). *J. Roy. Statist. Soc. Ser. B* **30** 31–66.

MORRIS, C. N., NORTON, E. C. and ZHOU, X. H. (1994). Parametric duration analysis of nursing home usage. In *Case Studies in Biometry* (N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest and J. Greenhouse, eds.) 231–248. Wiley, New York.

MURPHY, S. A. and VAN DER VAART, A. W. (1999). Observed information in semiparametric models. *Bernoulli* **5** 381–412.

MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–465.

NIELSEN, G. G., GILL, R. D., ANDERSEN, P. K. and SØRENSEN, T. I. A. (1992). A counting process approach to maximum likelihood estimator in frailty models. *Scand. J. Statist.* **19** 25–43.

PARNER, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.* **26** 183–214.

ROBINSON, P. M. (1988). The stochastic difference between econometrics and statistics. *Econometrica* **56** 531–548.

SINHA, D. (1998). Posterior likelihood methods for multivariate survival data. *Biometrics* **54** 1463–1474.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

TIBSHIRANI, R. J. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16** 385–395.

WAHBA, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13** 1378–1402.

DEPARTMENT OF STATISTICS
CHINESE UNIVERSITY OF HONG KONG
SHATIN, HONG KONG
E-MAIL: jfan@sta.cuhk.edu.hk

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802–2111
E-MAIL: rli@stat.psu.edu