Statistica Sinica Preprint No: SS-2016-0401R1	
Title	Asymptotic Behavior of Cox's Partial Likelihood and its Application to Variable Selection
Manuscript ID	SS-2016-0401
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0401
Complete List of Authors	Runze Li Jian-Jian Ren Guangren Yang and Ye Yu
Corresponding Author	Runze Li
E-mail	rzli@psu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

Asymptotic Behavior of Cox's Partial Likelihood and its Application to Variable Selection

Runze Li[1], Jian-Jian Ren[2], Guangren Yang[3] and Ye Yu[4]

[1]*Pennsylvania State University,* [2]*University of Maryland,*

[3]*Jinan University and* [4]*Wells Fargo Bank*

Abstract: For theoretical properties of variable selection procedures for Cox's model, we study the asymptotic behavior of partial likelihood for the Cox model. We find that the partial likelihood does not behave like an ordinary likelihood, whose sample average typically tends to its expected value, a finite number, in probability. Under some mild conditions, we prove that the sample average of partial likelihood tends to infinity at the rate of the logarithm of the sample size, in probability. We apply the asymptotic results on the partial likelihood to study tuning parameter selection for penalized partial likelihood. We find that the penalized partial likelihood with the generalized cross-validation (GCV) tuning parameter proposed in Fan and Li (2002) enjoys the model selection consistency property, despite the fact that GCV, AIC and C_p, equivalent in the context of linear regression models, are not model selection consistent. Our empirical studies via Monte Carlo simulation and a data example confirm our theoretical findings.

Key words and phrases: Akaike information criterion, Bayesian information criterion, LASSO penalized partial likelihood, SCAD, variable selection.

1. Introduction

The Cox model (Cox (1972)) has been the most popular model in the survival data

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

analysis during the past decades, and the partial likelihood (Cox (1975)) is perhaps the most commonly-used technique for analysis of right censored data. In practice, many risk factors and covariates are available for the initial analysis, thus an important task is to identify the significant risk factors and covariates. Variable selection is a useful technique in the analysis of survival data in the presence of many covariates. Classical variable selection criteria for linear regression models can be extended for the Cox model by replacing the log-likelihood by the log-partial likelihood (AIC (Akaike (1974)) and BIC (Schwarz (1978))). The LASSO (Tibshirani (1996)) variable selection technique has been extended for the Cox model (Tibshirani (1997); Zhang and Lu (2007); Zou (2008)). Nonconcave partial likelihood variable selection procedures have been developed for the Cox model (Fan and Li (2002); Bradic, Fan, and Jiang (2011)). To investigate the theoretical property of these procedures, we have to study the asymptotic behavior of the partial likelihood.

There has been little work on the asymptotic behavior of the partial likelihood, though the asymptotic properties of the partial likelihood estimator have been extensively studied (Tsiatis (1981); Andersen and Gill (1982); Takemi and Toshinari (1984)). Under mild regularity condition, the maximum partial likelihood estimator behaves the same as the ordinary maximum likelihood estimator of i.i.d. random samples in terms of asymptotic consistency, asymptotic normality and asymptotic efficiency. See, for example, Murphy and van der Vaart (2000). In this paper, we first study the asymptotic behavior of the partial likelihood, and prove that the '*sample average*' of partial likelihood diverges to infinity at a rate of the logarithm of the sample size. This clearly indicates that the Cox partial likelihood does not behave like an ordinary likelihood in that under mild regularity conditions, the sample average of the ordinary likelihood function converges to its expectation (a finite value) in

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

probability as the sample size tends to infinity.

With the aid of the asymptotic property of partial likelihood, we study the selection of regularization parameter in penalized partial likelihood for variable selection. Tibshirani (1997) proposed penalized partial likelihood with LASSO penalty for the Cox model. Fan and Li (2002) proposed the partial likelihood with the SCAD penalty for the Cox models, and showed that under certain regularity conditions, the resulting estimate enjoys the oracle property. Zhang and Lu (2007) and Zou (2008) further proposed adaptive LASSO for the Cox model to improve the SCAD procedure in terms of computational efficiency, while retaining the oracle property. The oracle property depends on the choice of the regularization parameter in penalized partial likelihood. It is well known that the regularization parameter controls the model complexity of the selected models, and plays a crucial role in these variable selection procedures. The issue of regularization parameter selection for penalized partial likelihood has not been systematically studied, in part because the asymptotic behavior of partial likelihood was not well understood. Wang, Li, and Tsai (2007) studied the selection of regularization parameter in the SCAD penalized least squares for linear regression models. They showed that with a positive probability, the *generalized cross-validation* (GCV, Craven and Wahba (1979)) selector yields an over-fitted model, and therefore this procedure does not enjoy the oracle property.

In this paper, we prove that the GCV selector for the SCAD method for the Cox model enjoys model selection consistency, in contrast to its model selection inconsistency in the least squares setting as demonstrated in Wang, Li, and Tsai (2007). Although GCV is equivalent to AIC and the C_p in the context of linear regression models, AIC and C_p yield an overfitted models with a positive probability, and thus are not model selection consistent.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

The rest of this paper is organized as follows. Section 2 studies the asymptotic behavior of the partial likelihood of the Cox model. We study the regularization parameter selection for the penalized partial likelihood in Section 3. Simulation study and a data example are presented in Section 4. Proofs are given in the Appendix.

2. Asymptotic Behavior of Cox's Partial Likelihood

Let T and $\mathbf{X} = (X_1, \cdots, X_d)^T$ be the survival time and associated d-dimensional vector of covariates, respectively. Consider the Cox proportional hazard regression model:

$$h(t \,|\, \mathbf{x}) = h_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta}), \tag{2.1}$$

where $\boldsymbol{\beta}$ is the regression coefficient vector, and $h(t \,|\, \mathbf{x})$ is the conditional hazard function of T given $\mathbf{X} = \mathbf{x}$ with $h_0(t)$ as an arbitrary baseline hazard function. Suppose that $(T_1, \mathbf{x}_1), \ldots, (T_n, \mathbf{x}_n)$ is a random sample of $(T, \mathbf{X})$, and the observed right censored survival data are as follows: $(V_1, \delta_1, \mathbf{x}_1)$, $\ldots$, $(V_n, \delta_n, \mathbf{x}_n)$, where $V_i = \min\{T_i, C_i\}, \delta_i = I\{T_i \leq C_i\}$, and C_i is the *right censoring variable* independent of T_i given $\mathbf{X} = \mathbf{x}_i$. Without loss of the generality, assume that there are no ties among observed continuous random variables V_i's. The log-partial likelihood function of the observed data is

$$\ell_c(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \delta_i \log \Big(\textstyle\sum_{j=1}^{n} I\{V_j \geq V_i\} \exp(\mathbf{x}_j^T \boldsymbol{\beta}) \Big). \tag{2.2}$$

(Cox (1975)). The goal is to study the asymptotic behavior of $\ell_c(\boldsymbol{\beta})$. We first illustrate the different behaviors of the log-partial likelihood and the likelihood of an i.i.d. sample by an example.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

Example 1. Suppose that we have an i.i.d. random sample $\{Y_1, \ldots, Y_n\}$ from a population with probability density/mass function $f(y;\theta)$, so $\ell(\theta) = \sum_{i=1}^{n} \log\{f(Y_i;\theta)\}$ is the log-likelihood function. By the weak law of large number, $n^{-1}\ell(\theta) \to E\log\{f(Y;\theta)\}$ in probability under mild regularity conditions. Furthermore, under mild regularity conditions, the maximum partial likelihood estimator, the maximizer of $\ell_c(\boldsymbol{\beta})$, behaves the same as the ordinary maximum likelihood estimator, the maximizer of $\ell(\theta)$, in terms of asymptotic consistency, asymptotic normality and asymptotic efficiency. See, for example, Murphy and van der Vaart (2000). Here, we numerically illustrate that

$$n^{-1}\ell_c(\boldsymbol{\beta}) \to \infty \quad \text{as} \quad n \to \infty. \tag{2.3}$$

We generated a random sample of size n from the proportional hazard model

$$h(t|x) = h_0(t)\exp(X\beta),$$

where $h_0(t) \equiv 1$, $\beta = 1$ and $X \sim N(0,1)$. The censoring variable C was generated from an exponential distribution with mean U. Therefore, the average censoring rate varies with different values of U. We list several values of U in Table 2.1 together with their corresponding average censoring rates, $1 - E\,I(T \le C) \widehat{=} 1 - \rho_1$, and take 10 different values of n ranging from $4(= 2^2)$ to $1024(= 2^{10})$. Figure 2.1 depicts the scatter plot of $\log(n)$ versus $-n^{-1}\ell_c$ based on a set of typical samples based on the different U listed in Table 2.1. Figure 2.1 clearly suggests that $-n^{-1}\ell_c$ increases at $\log(n)$ rate.

We next show that $-n^{-1}\ell_c(\boldsymbol{\beta})$ tends to infinite at the rate of $\log(n)$ using techniques

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

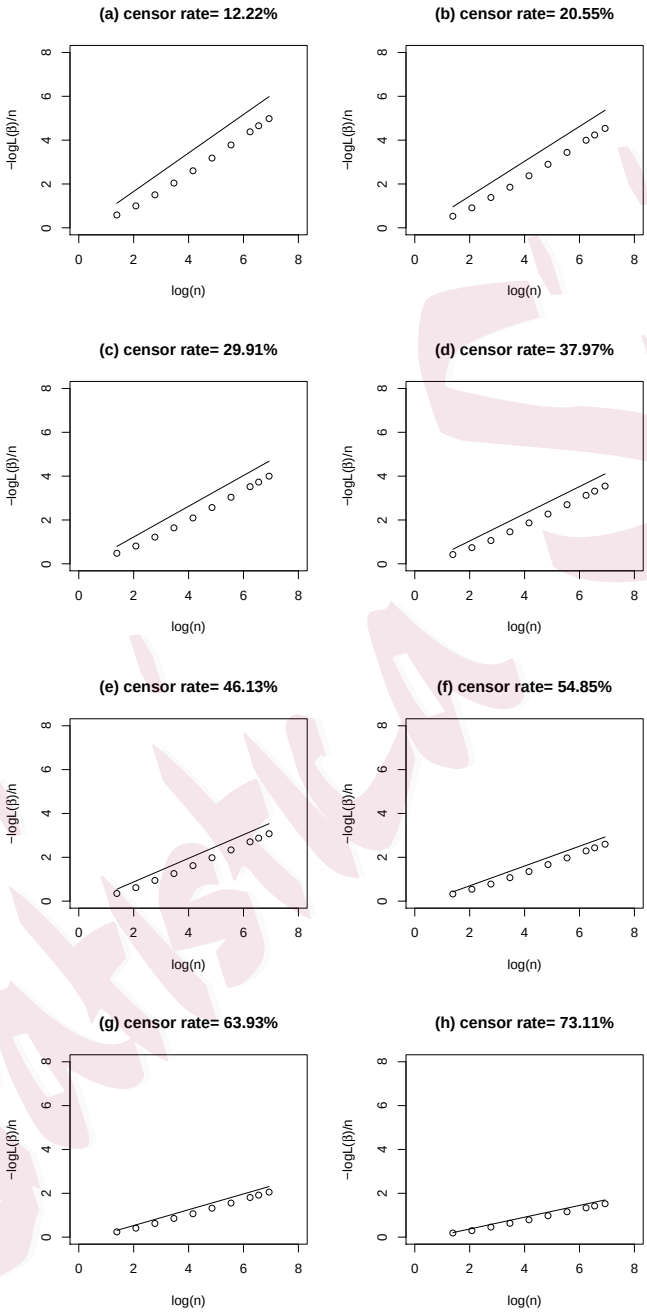

Figure 2.1: Plot of $\log(n)$ versus $-n^{-1}\ell_c$. 'o' is the scatter plot of $\log(n)$ versus $-n^{-1}\ell_c$ based on a typical simulated data set. The solid line in each plot is $\log(n)\widehat{\rho}_1 - \boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_0$ with $\boldsymbol{\beta} = 1$, where $\widehat{\rho}_1$ is an estimate of $EI\{T \le C\}$ and $\widehat{\boldsymbol{\mu}}_0$ is an estimate of $E\{I\{T \le C\}X\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

Table 2.1: Values of U and the corresponding Average Censoring Rates $(1-\rho_1)$ together with $\boldsymbol{\mu}_0 \widehat{=} \mathrm{E}\{(T \le C)X\}$.

U	10.00	5.00	2.75	1.80	1.20	0.80	0.50	0.30
$(1-\rho_1)$	0.1222	0.2055	0.2991	0.3797	0.4613	0.5485	0.6393	0.7311
$\boldsymbol{\mu}_0$	0.0968	0.1429	0.1775	0.1971	0.2007	0.2028	0.1921	0.1652

related to empirical processes. Let

$$G_n(v,\mathbf{x}) = n^{-1}\sum_{i=1}^{n} I\{V_i \le v, \mathbf{x}_i \le \mathbf{x}\}, \;\; H_n(v) = n^{-1}\sum_{i=1}^{n} I\{V_i \le v, \delta_i = 1\}, \tag{2.4}$$

with $G(v,\mathbf{x})$ and $H(v)$ as the limits of the empirical distribution functions $G_n(v,\mathbf{x})$ and $H_n(v)$, respectively. Take $\boldsymbol{\mu}_0 = E\{I\{T \le C\}\mathbf{X}\}$, $W(t) = \int\int_{v\ge t} \exp(\mathbf{x}^T\boldsymbol{\beta})\, dG(v,\mathbf{x})$, and $\rho_1 = EI\{T \le C\}$. The proof of the following theorem is given in Appendix A.

Theorem 1. *If $(V_i, \delta_i, \mathbf{x}_i)$, $i = 1, \cdots, n$, is a random sample from the Cox model (2.1) and the censoring time C_i is independent of T_i given $\mathbf{x}_i$, then the following statements hold.*

(a) If $\mathbf{X}$ has a finite bounded support, then

$$-n^{-1}\ell_c(\boldsymbol{\beta}) = \rho_1 \log n - \boldsymbol{\mu}_0^T\boldsymbol{\beta} + O_p(1), \quad as \quad n \to \infty. \tag{2.5}$$

(b) If $\mu_1 = \int_0^\infty \log W(t)\, dH(t)$ is well-defined, $E|X_j| < \infty$ for all $j = 1, \cdots, d$, and $0 < E\exp(\mathbf{X}^T\boldsymbol{\beta}) < \infty$,

$$-n^{-1}\ell_c(\boldsymbol{\beta}) = \rho_1 \log n - \boldsymbol{\mu}_0^T\boldsymbol{\beta} + \mu_1 + o_p(1). \tag{2.6}$$

When there is no censoring, it can be shown that $W(t) = f_T(t)/h_0(t)$ and $\mu_1 =$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

$\int_0^\infty \log[f_T(t)/h_0(t)]\, dF_T(t)$, where $f_T(t)$ and $F_T(t)$ are the probability density and cumulative distribution function of T in (2.1), respectively. Thus, the assumption about μ_1 holds for many distributions, such as the exponential distribution.

Remark. From the proof of this theorem, the leading term $\rho_1 \log(n)$ comes from $\log(n)(\frac{1}{n}\sum_{i=1}^n \delta_i)$, which does not depend on the regression coefficient $\boldsymbol{\beta}$ and does not affect the first and second order derivatives of the partial likelihood function. As the asymptotic normality of the maximum partial likelihood estimator relies on the first and second order derivatives, the divergent behavior of the partial likelihood function does not impact the asymptotic normality of the partial likelihood estimator. On other hand, the tuning parameter selector for penalized partial likelihood, studied in next section, depends on the partial likelihood function itself. As a result, the asymptotic behavior of the partial likelihood function directly affects the property of the tuning parameter selector.

3. Tuning parameter selector in penalized partial likelihood

Take the penalized partial likelihood to be

$$\ell_c(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.7}$$

where d is the dimension of $\boldsymbol{\beta}$, $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ (or more generally, λ_js). The penalized partial likelihood estimate of $\boldsymbol{\beta}$ maximizes (2.7) with respect to $\boldsymbol{\beta}$. Denote by $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$, and let $\boldsymbol{\beta}_0 = (\beta_{10}, \cdots, \beta_{d0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$. Without loss of generality, we take $\boldsymbol{\beta}_{20} = \mathbf{0}$ with all components of $\boldsymbol{\beta}_{10}$ nonzero. Under some regularity conditions, Fan and Li (2002) showed that the nonconcave penalized likelihood estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ possesses the oracle property: with probability tending to 1, for a

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

certain choice of $p_{\lambda_n}(\cdot)$, we have $\widehat{\boldsymbol{\beta}}_2 = 0$ and

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

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

The oracle property depends on the choice of the tuning parameter. Thus, the selection of tuning parameter is fundamental in the penalized likelihood procedure. Wang, Li, and Tsai (2007) studied the selection of the tuning parameter for penalized least squares for linear regression models. They showed that the GCV tuning parameter of Fan and Li (2001) cannot yield an oracle estimator. The issue of tuning parameter selection for the penalized partial likelihood has not been studied. Based on the asymptotic results about the partial likelihood, we show that the GCV tuning parameter selector for (2.7) possesses model selection consistency, in contrast to the model selection inconsistency of the GCV tuning parameter selector in the penalized least squares setting.

Let $\widehat{\boldsymbol{\beta}}_\lambda$ be the penalized partial likelihood estimator with tuning parameter λ. Define the GCV statistic to be

$$\mathrm{GCV}(\lambda) = \frac{-\ell_c(\widehat{\boldsymbol{\beta}}_\lambda)}{n\{1 - \mathrm{df}_\lambda/n\}^2}. \tag{2.8}$$

When $\widehat{\lambda}_{\mathrm{GCV}} = \mathrm{argmin}_\lambda\{\mathrm{GCV}(\lambda)\}$ is selected, where the degree of freedom df_λ is set to be the number of the nonzero penalized partial likelihood estimate corresponding to the tuning parameter λ. It can be shown that with probability tending to one, the effective number of parameters in Fan and Li (2002) is df_λ by using related techniques in Zhang, Li, and Tsai (2010).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

We define the corresponding AIC and BIC statistics as

$$\mathrm{AIC}(\lambda) = -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{df}_\lambda \tag{2.9}$$

$$\mathrm{BIC}(\lambda) = -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + \log(n)\mathrm{df}_\lambda, \tag{2.10}$$

with the AIC and BIC tuning parameter selectors

$$\widehat{\lambda}_{\mathrm{AIC}} = \mathrm{argmin}_\lambda\{\mathrm{AIC}(\lambda)\} \text{ and } \widehat{\lambda}_{\mathrm{BIC}} = \mathrm{argmin}_\lambda\{\mathrm{BIC}(\lambda)\}$$

When t lies in the neighborhood of 0, $(1-t)^{-2} \approx 1-2t$ so, when n is large enough,

$$2n\mathrm{GCV}(\lambda) \approx -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + 4(-\ell_c(\widehat{\boldsymbol{\beta}}_\lambda)/n)\mathrm{df}_\lambda.$$

If $-\ell_c(\widehat{\boldsymbol{\beta}}_\lambda)/(n\log(n)) \to EI\{T \le C\}$ as $n \to \infty$, then

$$2n\mathrm{GCV}(\lambda) \approx -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + 4\rho_1\log(n)\mathrm{df}_\lambda. \tag{2.11}$$

For $\rho_1 \ge 1/4$, the GCV tuning parameter can yield a sparser model than the one selected by the BIC-tuning parameter selector, as is seen in the simulation study in Section 4.

3.1. Definition and Notation

We first need to define the candidate models considered in model selection. Let $\bar{\alpha} = \{1, \cdots, d\}$ denote the label of predictors for the full model. Hence α, the subset of $\bar{\alpha}$, represents a candidate model including the predictors labelled by α. For each candidate model α, its model size and the corresponding coefficients are df_α and $\boldsymbol{\beta}_\alpha$. Therefore, each

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

tuning parameter λ determined in the penalty function results in a selected model α_λ with model size df_{α_λ} and the corresponding coefficients $\widehat{\boldsymbol{\beta}}_\lambda$. The collection of all candidate models is denoted by $\mathcal{A}$.

For any given model α, we are able to obtain its non-penalized estimates $\widehat{\boldsymbol{\beta}}^\star_\alpha$ by maximizing the corresponding partial likelihood $\ell_c(\boldsymbol{\beta})$. Similarly, for any selected model α_λ obtained from penalized partial likelihood with given λ, we are able to obtain the corresponding non-penalized estimates $\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}$.

To study the asymptotic behaviors of the tuning parameter selectors for penalized partial likelihood, we define a general tuning parameter selector

$$\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}) = -2\ell_c(\widehat{\boldsymbol{\beta}}) + \kappa_n \mathrm{df}_{\widehat{\boldsymbol{\beta}}}, \tag{2.12}$$

where $\widehat{\boldsymbol{\beta}}$ is the parameter estimator and $\mathrm{df}_{\widehat{\boldsymbol{\beta}}}$ is the corresponding degree freedom associated with $\widehat{\boldsymbol{\beta}}$. Here κ_n is a positive number that denotes different variable selection criterion. When $\kappa_n = 2$, GIC_{κ_n} is the AIC at (2.9), and when $\kappa_n = \log(n)$, GIC_{κ_n} is the BIC at (2.10).

3.2. Theoretical Property

In this section, we assume that the set of candidate models contain the unique true model and that the number of parameters in the full model is finite. Assume that the coefficients of the unique true model α_0 in $\mathcal{A}$ are nonzero. Therefore, any candidate model $\alpha \not\supseteq \alpha_0$ is an underfitted model while any model $\alpha \supset \alpha_0$ is an overfitted model. We partition the tuning parameters into

$$\Omega_- = \{\lambda : \alpha_\lambda \not\supseteq \alpha_0\}, \quad \Omega_0 = \{\lambda : \alpha_\lambda = \alpha_0\} \quad \text{and} \quad \Omega_+ = \{\lambda : \alpha_\lambda \supset \alpha_0\}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

We need the following conditions.

(E1) $\lambda_{\max}$ depends on n and satisfies $\lambda_{\max} \to 0$ as $n \to \infty$.

(E2) There exits a constant m such that the penalty $p_\lambda(\xi)$ satisfies $p'_\lambda(\xi) = 0$ for $\xi > m\lambda$.

(E3) If $\lambda_n \to 0$ as $n \to \infty$, then the penalty function satisfies

$$\liminf_{n\to\infty} \liminf_{\xi\downarrow 0} \sqrt{n} p'_\lambda(\xi) \to \infty.$$

(E4) For any candidate model $\alpha \in \mathcal{A}$, there exits $c_\alpha > 0$ such that $-n^{-1}\ell_c(\widehat{\boldsymbol{\beta}}^\star_\alpha) - \log(n)\rho_1 \to c_\alpha$. In addition, for any underfitted model $\alpha \not\supseteq \alpha_0$, $c_\alpha > c_{\alpha_0}$.

Conditions (E1)-(E3) are conditions on the penalty while condition (E4) is the technical condition needed to investigate the asymptotic properties of the tuning parameter selectors for penalized partial likelihood. Condition (E1) indicates a smaller tuning parameter is required if the sample size is large; (E2) implies that the penalty is chosen to have an asymptotic unbiased estimator; (E3) is used to study the oracle property of the penalized estimator; (E4) assures that the underfitted model yields a larger model deviance than that of the true model.

Theorem 2. *Suppose that the partial likelihood function of the Cox's model satisfies Conditions (A)-(D) in Fan and Li (2002) and that Conditions (E1)-(E4) hold.*

(A) If there exits a positive constant M such that $\kappa_n < M$, then the tuning parameter $\widehat{\lambda}$ obtained by minimizing $GIC_{\kappa_n}(\lambda)$ satisfies $P\{\widehat{\lambda} \in \Omega_-\} \to 0$ and $P\{\widehat{\lambda} \in \Omega_+\} > 0$.

(B) If $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$, then the tuning parameter $\widehat{\lambda}$ obtained by minimizing $GIC_{\kappa_n}(\lambda)$ satisfies $P\{\alpha_{\widehat{\lambda}} = \alpha_0\} \to 1$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

(C) If $\rho_1 > 0$, then the tuning parameter $\widehat{\lambda}$ obtained by minimizing the GCV score defined in (2.8) satisfies $P\{\alpha_{\widehat{\lambda}} = \alpha_0\} \to 1$.

The proof of Theorem 2 is given in the supplement (Li, Ren, Yang and Yu (2016)).

Here, Theorem 2(A) implies that the GIC_{κ_n} selector with bounded κ_n tends to overfit without considering which penalty function is used, while Theorem 2(B) indicates that the GIC_{κ_n} selector with diverging κ_n enables us to identify the true model consistently. Thus, the penalized partial likelihood with diverging κ_n possesses the oracle property. Theorem 2(C) implies that the penalized partial likelihood estimator with the GCV selector also possesses the oracle property. This is quite different from penalized least squares for the linear regression model; as shown in Wang, Li and Tsai (2007), the GCV selector for the penalized least squares with linear model results in an overfitted model with positive probability.

4. Numerical Results

We assessed the finite sample performance of proposed procedures. Since there exist various comparisons among penalized partial likelihood with different penalties such as the LASSO and SCAD. In our simulation studies, we focused on comparisons among different tuning parameter selectors for penalized partial likelihood with the SCAD penalty. For simplicity, we refer to the SCAD penalized partial likelihood with $\kappa_n = 2$ and $\log(n)$ in GIC_{κ_n} tuning parameter selector as SCAD-AIC and SCAD-BIC, respectively. Similarly we refer to the SCAD method with the GCV as SCAD-GCV. The best subset selection with AIC and BIC criteria for the Cox model are denoted by AIC and BIC in this section, respectively. In our simulation, we employed the local linear algorithm (LLA, Zou and Li (2008)) to compute the parameter estimates of the SCAD penalized partial likelihood function.

Example 4.1. We adapted the model structure in Fan and Li (2002) to generate the data

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

with sample sizes $n = 100, 200$, and 400 from the Cox model with hazard function

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

where $h_0(t) \equiv 1$, $\boldsymbol{\beta} = (0.8, 0, 0, 1, 0, 0, 0, 0, 0, 0.6, 0, 0)^T$, and $\mathbf{x}$ had a 12-dimensional normal distribution, with the correlation between x_i and x_j as $0.5^{|i-j|}$. Accordingly, $\mu(\mathbf{x}^T\boldsymbol{\beta}) = \exp(-\mathbf{x}^T\boldsymbol{\beta})$. The censoring distribution was exponential with mean $U \exp(-\mathbf{x}^T\boldsymbol{\beta})$, where U was sampled from a uniform distribution over $[1, 3]$. Consequently, the average censoring percentage was 35%. We include the case with no censoring as a benchmark. For each scenario, we conducted 1000 simulations.

Table 2.2: Simulation results for the Cox model (**No Censoring**)

		MRME	Zeros		Under	Exact	Over Fitted (%)				
n	Method	(%)	C	IC	(%)	(%)	1	2	3	4	≥ 5
100	SCAD-AIC	45.75	7.255	0.001	0.1	37.5	15.3	16.1	13.1	8.5	9.4
	SCAD-BIC	20.90	8.576	0.003	0.3	74.0	15.7	5.2	3.8	0.9	0.1
	SCAD-GCV	17.29	8.940	0.059	5.6	89.2	4.9	0.3	0.0	0.0	0.0
	AIC	52.52	7.349	0.001	0.1	20.1	29.4	26.9	15.3	6.0	2.2
	BIC	25.68	8.666	0.004	0.4	72.5	22.6	3.4	1.1	0.0	0.0
	Oracle	15.73	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0
200	SCAD-AIC	58.53	7.591	0.000	0.0	46.6	15.9	13.8	8.5	7.3	7.9
	SCAD-BIC	36.33	8.867	0.000	0.0	90.1	7.3	1.8	0.8	0.0	0.0
	SCAD-GCV	33.96	8.995	0.003	0.3	99.2	0.5	0.0	0.0	0.0	0.0
	AIC	66.37	7.506	0.000	0.0	23.1	32.2	24.8	13.8	4.5	1.6
	BIC	41.89	8.781	0.000	0.0	81.2	16.1	2.3	0.4	0.0	0.0
	Oracle	33.95	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0
400	SCAD-AIC	68.14	7.553	0.000	0.0	45.1	14.7	15.4	9.3	9.1	6.4
	SCAD-BIC	44.10	8.936	0.000	0.0	94.7	4.4	0.7	0.2	0.0	0.0
	SCAD-GCV	42.33	8.999	0.000	0.0	99.9	0.1	0.0	0.0	0.0	0.0
	AIC	74.71	7.530	0.000	0.0	22.6	33.4	25.7	12.2	5.0	1.1
	BIC	47.38	8.875	0.000	0.0	88.6	10.5	0.7	0.2	0.0	0.0
	Oracle	42.30	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0

To assess finite sample performance, we report the percentage of models correctly fitted, underfitted, and overfitted with 1, 2, 3, 4, 5 or more parameters by five variable selection

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

Table 2.3: Simulation results for the Cox model (**35% Censoring**)

		MRME	Zeros		Under	Exact	Over Fitted (%)				
n	Method	(%)	C	IC	(%)	(%)	1	2	3	4	≥ 5
100	SCAD-AIC	42.43	7.235	0.012	1.2	33.4	18.8	16.6	12.0	8.5	9.5
	SCAD-BIC	21.42	8.491	0.060	5.8	63.4	19.7	7.3	2.4	1.1	0.3
	SCAD-GCV	19.04	8.800	0.153	13.6	71.6	12.3	2.1	0.3	0.1	0.0
	AIC	50.03	7.370	0.016	1.6	20.4	30.0	25.9	13.6	6.5	2.0
	BIC	23.45	8.648	0.036	3.6	68.8	23.7	3.5	0.4	0.0	0.0
	Oracle	14.35	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0
200	SCAD-AIC	59.24	7.535	0.000	0.0	42.3	19.5	13.2	10.2	7.7	7.1
	SCAD-BIC	35.53	8.841	0.000	0.0	87.4	9.8	2.3	0.5	0.0	0.0
	SCAD-GCV	32.48	8.963	0.006	0.6	95.9	3.3	0.2	0.0	0.0	0.0
	AIC	64.64	7.513	0.000	0.0	22.8	35.5	21.5	12.1	6.9	1.2
	BIC	37.90	8.830	0.000	0.0	84.8	13.5	1.6	0.1	0.0	0.0
	Oracle	31.45	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0
400	SCAD-AIC	69.31	7.552	0.000	0.0	41.5	19.5	14.7	10.6	7.4	6.3
	SCAD-BIC	45.07	8.920	0.000	0.0	93.2	5.7	1.0	0.1	0.0	0.0
	SCAD-GCV	42.75	8.993	0.000	0.0	99.4	0.5	0.1	0.0	0.0	0.0
	AIC	73.64	7.547	0.000	0.0	23.8	33.7	24.7	11.6	4.5	1.7
	BIC	48.85	8.856	0.000	0.0	86.8	12.0	1.2	0.0	0.0	0.0
	Oracle	43.47	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0

procedures, as well as the simulated data fitted with the true model over 1000 simulations. We report the average number of zero coefficients that were correctly (C) and incorrectly (IC) identified in the selected models over 1000 simulations. To compare model fittings, we calculated the model error for the new observation $(V, \delta, \mathbf{x})$,

$$ME(\widehat{\boldsymbol{\beta}}) = E_{\mathbf{X}}\{\mu(\mathbf{x}^T\boldsymbol{\beta}) - \mu(\mathbf{x}^T\widehat{\boldsymbol{\beta}})\}^2,$$

where the expectation is taken with respect to the new observed covariate vector $\mathbf{x}$, and $\mu(\mathbf{x}^T\boldsymbol{\beta}) = E(T|\mathbf{x}, \boldsymbol{\beta})$. We report the median of the relative model error (MRME) over 1000 simulations, where the relative model error is defined as $RME = ME/ME_{full}$, and ME_{full} is the model error calculated by fitting the data with the full model.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

In Fan and Li (2002), it was shown that

$$ME(\widehat{\boldsymbol{\beta}}) = E_{\mathbf{X}}\{\mu(\mathbf{x}^T\boldsymbol{\beta}) - \mu(\mathbf{x}^T\widehat{\boldsymbol{\beta}})\}^2 = E_{\mathbf{X}}\{\exp(-\mathbf{x}^T\boldsymbol{\beta}) - \exp(-\mathbf{x}^T\widehat{\boldsymbol{\beta}})\}^2.$$

By using the moment generating function of the multinormal distribution, we can simplify this to

$$ME(\widehat{\boldsymbol{\beta}}) = \exp(2\widehat{\boldsymbol{\beta}}^T\Sigma\widehat{\boldsymbol{\beta}}) + \exp(2\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}) - 2\exp\{\frac{1}{2}(\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta})^T\Sigma(\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta})\},$$

where Σ is the covariance matrix of $\mathbf{x}$. We use this formula to calculate model errors for our simulations.

Table 2.2 gives the results for the uncensored case, and shows that the MRME of SCAD-BIC/GCV is smaller than that of SCAD-AIC. As the sample size increases, the MRME of SCAD-BIC/GCV approaches that of the oracle estimator, whereas the MRME of SCAD-AIC remains at the same level. Interestingly, SCAD-BIC and SCAD-AIC have smaller MRME than that of the best subset selection with BIC and AIC, respectively.

Table 2.2 also shows that SCAD-BIC/GCV has a higher probability of correctly estimating the true zero coefficients to zero than does SCAD-AIC. However, SCAD-BIC/GCV was more prone than SCAD-AIC to incorrectly set the three nonzero coefficients to zero when the sample size was small, and SCAD-GCV was more aggressive than SCAD-BIC with larger values in "IC" columns. In addition, SCAD-BIC/GCV had a much higher probability of correctly identifying the true model.

For the censored case, Table 2.3 shows findings similar to those presented in Table 2.2. Accordingly, SCAD-BIC/GCV was superior to SCAD-AIC in both identifying the true

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

model, and in reducing the model error and complexity. When the data was 35% censored, all methods declined slightly in their efficacy, while the relative performance of SCAD-BIC/GCV versus SCAD-AIC remained the same as in the uncensored case. This is consistent with our theoretical analysis in Section 3.

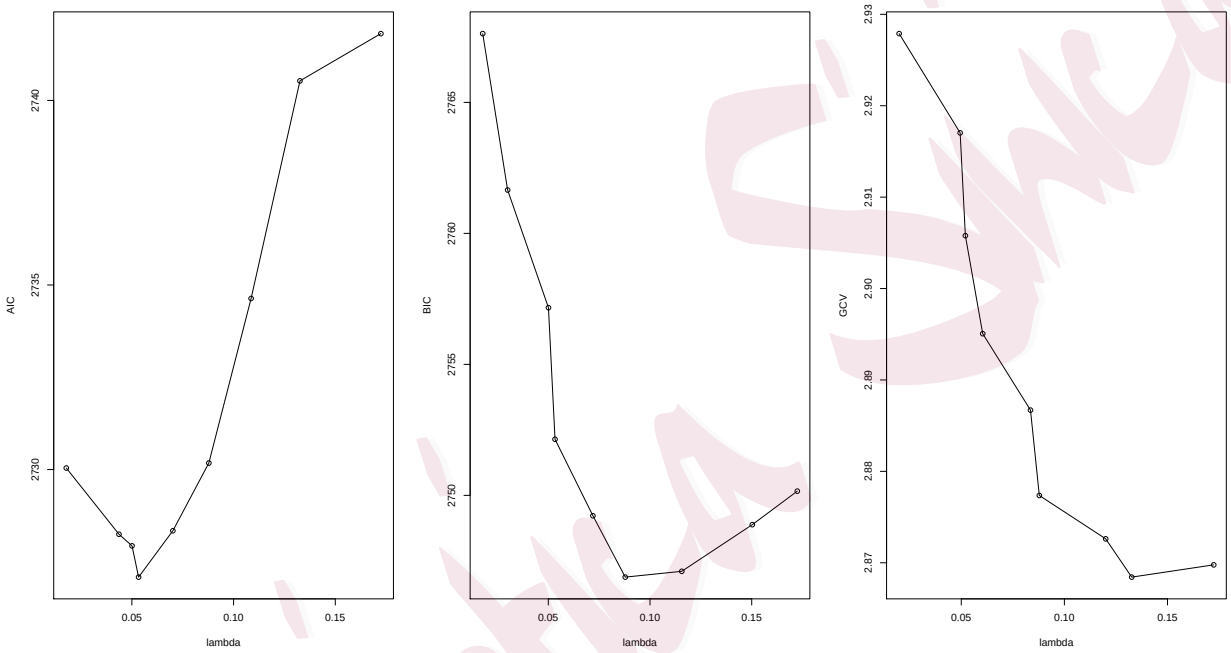

Figure 2.2: The left panel is the GIC scores with $\kappa_n = 2$ versus λ, the middle panel is the GIC score with $\log(n)$ versus λ, and the right panel is the GCV scores versus λ.

Example 4.2. (Heart attack data) We applied the proposed regularization parameter selection procedures to the heart attack data set used in Hosmer and Lemeshow (1999). The data were collected in the Worcester Heart Attack Study which describes trends over time in survival rates following hospital admission for acute myocardial infarction. The total length of follow-up on the admission of 481 hospital patients was recorded for years 1975, 1978, 1981, 1984, 1986, and 1988. Among those patients, 249 died and the rest were censored at the rate of 48%.

To model survival time, Hosmer and Lemeshow (1999) suggested fitting the Cox propor-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

Table 2.4: Estimates and Standard Errors for Heart Attack Data

	MPLE	SCAD-AIC	SCAD-BIC	SCAD-GCV
age (x_1)	0.60(0.13)	0.56(0.09)	0.43(0.07)	0.41(0.05)
cpk (x_2)	0.03(0.14)	0(-)	0(-)	0(-)
sex (x_3)	0.17(0.14)	0(-)	0(-)	0(-)
chf (x_4)	0.80(0.14)	0.80(0.13)	0.80(0.14)	0.82(0.13)
miord (x_5)	0.42(0.14)	0.43(0.13)	0.41(0.13)	0(-)
age*sex(x_6)	-0.29(0.14)	-0.22(0.13)	0(-)	0(-)
age*chf (x_7)	-0.07(0.15)	0(-)	0(-)	0(-)
age*miord (x_8)	0.03(0.15)	0(-)	0(-)	0(-)
cpk*sex (x_9)	-0.16(0.16)	0(-)	0(-)	0(-)
cpk*chf (x_{10})	0.19(0.15)	0.19(0.09)	0(-)	0(-)
cpk*miord (x_{11})	0.29(0.15)	0.25(0.12)	0.21(0.05)	0(-)

tional hazards model with five explanatory variables: x_1-*age*; x_2-*cpk* (peak cardiac enzymes in international units); x_3-*sex* (male=0 and female=1); x_4-*chf* (left heart failure complications, yes=1 and no=0); x_5-*miord* (MI order, first=0 and recurrent=1). In addition to these variables, we included the six interactions between the two continuous variables (*age* and *cpk*) and the three indicator variables (*sex, chf*, and *miord*). Thus, there were 11 variables in our full model. We applied the penalized partial likelihood approach. The resulting regularization parameters selected by SCAD-AIC, SCAD-BIC, and SCAD-GCV were 0.0533, 0.0878, and 0.1326, respectively. The corresponding tuning parameters selector curves are depicted in Figure 2.2.

Table 2.4 presents the maximum partial penalized likelihood estimates (MPLE) from the full model as well as the SCAD-AIC/BIC/GCV parameter estimates, together with their standard errors. The full model contained six insignificant variables (x_2, x_3, and x_7 to x_{10}) at level 0.05, SCAD-AIC included two insignificant variables (x_6 and x_{10}) at level 0.05. In contrast, the four variables x_1, x_4, x_5, and x_{11}, selected by SCAD-BIC were significant at level 0.05. For this data set, SCAD-GCV looks to be overly aggressive in that it excludes

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

x_5, and x_{11}.

Based on Table 2.4, the p-values of the partial likelihood ratio test for examining the SCAD-AIC, SCAD-BIC, and SCAD-GCV model versus the full model are 0.6752, 0.1749, and 0.0034, respectively. Consequently, there is no evidence of lack of fit in the SCAD-BIC model. The SCAD-GCV model may be too aggressive, consistent with our simulation results that GCV tends to be underfitted when the sample size is not large enough.

5. A tribute to Peter Hall

Professor Peter Hall made wide ranging and ground-breaking contributions to many statistical fields and played major leadership roles throughout the statistical profession. He was a true scholar, and a mentor and friend of many of us. We grieve his loss.

Runze Li (RL) had the great fortune to learn from Peter and interact with him directly when they jointly served as Editors of the *Annals of Statistics* from 2013 to 2015. As an eminent scientist, Peter was an extremely kind, modest and optimistic person. Peter was always super fast, and handled whatever came to him promptly. His speed was unbeatable. Once, RL was asked to review a grant proposal by an international grant agency within a tight deadline. When RL sent back his report the next day, he was told that Peter's report had already been received.

Professor Peter Hall had a huge influence on RL's research on variable selection and feature screening, although he never collaborated with Peter on a paper. Many of RL's works were inspired by Peter's ideas. For example, Hall and Miller (2009) proposed using generalized correlation to conduct feature screening and the use of the bootstrap to quantify the uncertainty of feature ranking. Motivated by this work, Li, Zhong, and Zhu (2012) proposed using distance correlation for feature screening.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

Professor Peter Hall will be remembered forever as a legendary statistician, a great scholar, beloved colleague, mentor and friend, and his work will continue to have a far-reaching impact on statistical methodology and theory.

Supplemental Materials

The proof of Theorem 2 is in the supplemental materials of this paper.

Acknowledgement

The authors would like to thank the editors Professors Raymond J. Carroll and Qiwei Yao for organizing this special issue. Runze Li is grateful to the editors for their invitation and their constructive comments on an earlier version of this paper. Li's research is supported by National Institute on Drug Abuse grants P50 DA039838, P50 DA036107, and R01 DA039854, National Science Foundation (NSF) grant DMS 1512422, and National Library of Medicine grant T32 LM012415. His research was also partially supported by National Nature Science Foundation of China grants 11690014 and 11690015. Ren's research is supported by NSF grants DMS 0905772, DMS 1232424, and DMS 1407461. Yang's research was supported by the NNSFC grant 11471086, the National Social Science Foundation of China grant 16BTJ032, the Fundamental Research Funds for the Central Universities 15JNQM019 and 21615452, the National Statistical Scientific Research Center Projects 2015LD02, the China Scholarship Council 201506785010 and Science and Technology Program of Guangzhou 2016201604030074. All authors equally contributed to this paper and are listed in alphabetic order. Guangren Yang is the corresponding author. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, the NIDA, or the NIH.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

Appendices

Appendix A: Proof of Theorem 1

Without loss of generality, assume that there are no ties among V_i's in the observed data, and that

$$V_1 < V_2 < \ldots < V_n. \tag{A.1}$$

This simplifies $n^{-1}\ell_c(\boldsymbol{\beta})$ to

$$n^{-1}\ell_c(\boldsymbol{\beta}) = n^{-1}\boldsymbol{\beta}^T \sum_{i=1}^{n} \delta_i \mathbf{x}_i - n^{-1} \sum_{i=1}^{n} \delta_i \log\Big(\exp(\mathbf{x}_i^T\boldsymbol{\beta}) + \cdots + \exp(\mathbf{x}_n^T\boldsymbol{\beta})\Big). \tag{A.2}$$

It follows by the Weak Law of Large Numbers (WLLN) that $(n^{-1}\sum_{i=1}^{n}\delta_i\mathbf{x}_i)^T\boldsymbol{\beta} = \boldsymbol{\mu}_0^T\boldsymbol{\beta} + o_P(1)$. Let

$$R_n = n^{-1} \sum_{i=1}^{n} \delta_i \log\Big(\exp(\mathbf{x}_i^T\boldsymbol{\beta}) + \cdots + \exp(\mathbf{x}_n^T\boldsymbol{\beta})\Big). \tag{A.3}$$

Thus,

$$-n^{-1}\ell_c(\boldsymbol{\beta}) = -\boldsymbol{\mu}_0^T\boldsymbol{\beta} + R_n + o_P(1), \tag{A.4}$$

From (A.1), we have

$$\begin{aligned}
&\exp(\mathbf{x}_i^T\boldsymbol{\beta}) + \cdots + \exp(\mathbf{x}_n^T\boldsymbol{\beta}) = \sum_{j=1}^{n} I\{V_j \geq V_i\} e^{\mathbf{x}_j^T\boldsymbol{\beta}} \\
&= \int\int I\{v \geq V_i\} \exp(\mathbf{x}^T\boldsymbol{\beta})\, d\Big\{ \sum_{j=1}^{n} I\{V_j \leq v, \mathbf{x}_j \leq \mathbf{x}\}\Big\} \\
&= \int\int_{v \geq V_i} \exp(\mathbf{x}^T\boldsymbol{\beta})\, d\Big\{ \sum_{j=1}^{n} I\{V_j \leq v, \mathbf{x}_j \leq \mathbf{x}\}\Big\} = nW_n(V_i),
\end{aligned} \tag{A.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

where $W_n(t) = \int\int_{v\geq t} \exp(\mathbf{x}^T\boldsymbol{\beta})\, dG_n(v,\mathbf{x})$ with $G_n(v,\mathbf{x})$ given in (2.4). Here δ_i is a binary random variable, $n^{-1}\sum_{i=1}^n \delta_i = \rho_1 + O_P(1/\sqrt{n})$. With $A_n = \int_0^\infty \log[W_n(t)]\, dH_n(t)$, it follows that

$$\begin{aligned} R_n &= n^{-1}\sum_{i=1}^n \delta_i \log\Big(nW_n(V_i)\Big) = n^{-1}\int_0^\infty \log\Big(nW_n(t)\Big)\, d\Big\{\sum_{i=1}^n \delta_i I\{V_i\leq t\}\Big\} \\ &= \int_0^\infty \Big\{\log n + \log\Big(W_n(t)\Big)\Big\}\, dH_n(t) = \log n\Big(H_n(\infty) - H_n(0)\Big) + A_n \\ &= \log n\Big(n^{-1}\sum_{i=1}^n \delta_i\Big) + A_n = \rho_1 \log n + \log n\Big(n^{-1}\sum_{i=1}^n \delta_i - \rho_1\Big) + A_n \\ &= \rho_1 \log n + O_P(n^{-1/2}\log n) + A_n = \rho_1 \log n + A_n + o_P(1), \end{aligned} \tag{A.6}$$

To prove Part (a), we next deal with A_n. Since

$$(n-i+1)\min_{i\leq j\leq n} \exp(\mathbf{x}_j^T\boldsymbol{\beta}) \leq \sum_{j=i}^n \exp(\mathbf{x}_j^T\boldsymbol{\beta}) \leq (n-i+1)\max_{i\leq j\leq n} \exp(\mathbf{x}_j^T\boldsymbol{\beta})$$

and $\mathbf{X}$ has a finite bounded support, it follows

$$\begin{aligned} A_n &= n^{-1}\sum_{i=1}^n \delta_i \log\Big(W_n(V_i)\Big) = n^{-1}\sum_{i=1}^n \delta_i \log\Big(\frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta}) + \cdots + \exp(\mathbf{x}_n^T\boldsymbol{\beta})}{n}\Big) \\ &= n^{-1}O_P\Big(\sum_{i=1}^n \log\big(\frac{n-i+1}{n}\big)\Big) = n^{-1}O_P(\log(n!/n^n)) = O_P(1). \end{aligned} \tag{A.7}$$

The last equality is due to Sterling's formula. and this completes the proof of (a).

For Part (b), it suffices to show that

$$A_n \xrightarrow{P} \mu_1, \qquad \text{as } n\to\infty. \tag{A.8}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

From (2.4), we know that $H_n(v)$ is the empirical process of a random sample of V_i's with $\delta_i = 1$. Thus, $\|H_n - H\| = \sup_v |H_n(v) - H(v)| = O_p(n^{-1/2})$ by the DWK inequality (van der Vaart (1998)) since $EI\{\delta_i = 1\} = \rho_1 > 0$. Hence, from (2.4), (A.5), and integration by parts, we have

$$\begin{aligned} A_n &= \int_0^{V_n} \log[W_n(t)]\, dH_n(t) = B_n + \int_0^{V_n} \log[W_n(t)]\, d[H_n(t) - H(t)] \\ &= B_n + [H_n(t) - H(t)] \log[W_n(t)]\Big|_0^{V_n} - \int_0^{V_n} [H_n(t) - H(t)] d\{\log[W_n(t)]\} \\ &= B_n + [H_n(V_n) - H(V_n)] \log[W_n(V_n)] - \int_0^{V_n} [H_n(t) - H(t)] d\{\log[W_n(t)]\} \\ &= B_n + [H_n(V_n) - H(V_n)] \log\{\exp(\mathbf{x}_n^T\boldsymbol{\beta})/n\} - \int_0^{V_n} [H_n(t) - H(t)] d\{\log[W_n(t)]\} \\ &= B_n + O_p\Big(\frac{\log n}{\sqrt{n}}\Big) - \int_0^{V_n} [H_n(t) - H(t)] d\{\log[W_n(t)]\}, \end{aligned} \tag{A.9}$$

where $B_n = \int_0^{V_n} \log[W_n(t)] dH(t)$ by using the fact that $\mathbf{x}_n^T\boldsymbol{\beta} = O_P(1)$, since $E(|X_j|) < \infty$ by the assumption on $E|X_j| < \infty$ for all $j = 1, \cdots, p$. From (2.4) and (A.5), we have

$$\begin{aligned} \Big| \int_0^{V_n} [H_n(t) - H(t)] d\{\log[W_n(t)]\} \Big| &\le \|H_n - H\| \times |\log W_n(V_n) - \log W_n(0)| \\ &= O_p(n^{-1/2}) \log\Big(\frac{\exp(\mathbf{x}_1^T\boldsymbol{\beta}) + \cdots + \exp(\mathbf{x}_n^T\boldsymbol{\beta})}{\exp(\mathbf{x}_n^T\boldsymbol{\beta})}\Big). \end{aligned}$$

By the assumption in Part (b) and the WLLN, $\frac{1}{n}\sum_{i=1}^n \exp(\mathbf{x}_i^T\boldsymbol{\beta}) \xrightarrow{P} E\exp\{\mathbf{X}^T\boldsymbol{\beta}\}$. This implies that $\log\{\sum_{i=1}^n \exp(\mathbf{x}_i^T\boldsymbol{\beta})\} - \log(n) = O_P(1)$. Furthermore, $\log\{\exp(\mathbf{x}_n^T\boldsymbol{\beta})\} = \mathbf{x}_n^T\boldsymbol{\beta} = O_P(1)$. Thus,

$$\log\Big(\frac{\exp(\mathbf{x}_1^T\boldsymbol{\beta}) + \cdots + \exp(\mathbf{x}_n^T\boldsymbol{\beta})}{\exp(\mathbf{x}_n^T\boldsymbol{\beta})}\Big) = O_P\{\log(n)\}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

It then follows that

$$\Big| \int_0^{V_n} [H_n(t) - H(t)] d\{\log[W_n(t)]\} \Big| = O_p\Big(\frac{\log n}{\sqrt{n}}\Big). \tag{A.10}$$

Therefore, (A.8) follows from (A.9)-(A.10), the assumption about μ_1, and the Dominated Convergence Theore.Thus,

$$B_n \xrightarrow{P} \mu_1, \qquad \text{as } n \to \infty. \tag{A.11}$$

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, **10**, 1100–1120.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Annals of Statistics*, **39**, 3092–3120.

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

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. *Numerische Mathematik,* **31**, 377–403.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

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

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74–99.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, **18**, 533 - 550.

Hosmer, D. W. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. John Wiley & Sons Inc., New York, NY.

Li, R. Ren, J.-J., Yang, G. and Yu, Y. (2016). Supplement to "Asymptotic behavior of Cox's partial likelihood and its application to variable selection".

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of American Statistical Association*, **107**, 1129 - 1139.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of American Statistical Association*, **95**, 449–465.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **19**, 461–464.

Takemi, Y. and Toshinari K. (1984). Maximum full and partial likelihood estimators in the proportional hazard model. *Annals of the Institute of Statistical Mathematics*, **36**, 363–373.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

Tibshirani, R. (1996). Regression shrinkage and selection via LASSO, *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

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

Tsiatis, A.A. (1981). A large sample study of Cox's regression model. *The Annals of Statistics*, **9**, 93–108.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge U. Press.

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

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of American Statistical Association*, **105**, 312–323.

Zhang, H. and Lu, W. (2007). Adaptive LASSO for Cox's proportional hazards model. *Biometrika*, **94**, 691–703.

Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*, **95**, 241–247.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics*, **36**, 1509–1566.

Runze Li

Department of Statistics and The Methodology Center,

The Pennsylvania State University, University Park, PA 16802.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

E-mail: rzli@psu.edu

Jian-Jian Ren

Department of Mathematics, University of Maryland, College Park, MD 20742.

Email: jjren@umd.edu

Guangren Yang

School of Economics, Jinan University, Guangzhou, P.R. China 510632.

Email: tygr@jnu.edu.cn.

Ye Yu

Quantitative Associate, Wells Fargo Bank

550 California Street, San Francisco, CA 94104

E-mail: Ye.Yu@wellsfargo.com