Supplement to "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*

This supplement consists of the proof of Theorem 2 in the main text.

Proof of Theorem 2. To prove Theorem 2, we show the following two lemmas. Theorem 2(A) and 2(B) follow Lemma 1 and 2 respectively.

Lemma 1. *Suppose that the partial likelihood function of the Cox model satisfies Conditions (A)-(D) in Fan and Li (2002). Assume that there exits a positive constant M such that $\kappa_n < M$. Then under Condition (E4),we have*

$$P\{\inf_{\lambda\in\Omega_-} GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}^\star_{\bar{\alpha}})\} \to 1 \quad as\ n\to\infty, \tag{S.1}$$

$$\liminf_{n\to\infty} P\{\inf_{\lambda\in\Omega_0} GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}^\star_{\bar{\alpha}})\} \geq \pi. \tag{S.2}$$

Proof. Recall that for any given λ, we can obtain a selected model α_λ by penalized variable selection. And based on this selected model α_λ, we are able to obtain its corresponding non-penalized estimates $\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}$ by maximizing the corresponding partial likelihood. Then

$$\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}) \geq \ell_c(\widehat{\boldsymbol{\beta}}_\lambda), \tag{S.3}$$

and $-2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + \kappa_n \mathrm{df}_\lambda > -2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda})$ Thus,

$$\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > -2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}). \tag{S.4}$$

Subtract $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})$ from both size of (S.4), we can obtain that

$$\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}}) > -2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_\lambda}) - \{-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}}) + \kappa_n \mathrm{df}_{\bar{\alpha}}\}.$$

For any $\lambda \in \Omega_- = \{\lambda : \alpha \not\supseteq \alpha_0\}$, we can take $\inf_{\lambda\in\Omega_-}$ over $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda})$. Under Condition (E4) and $\kappa_n < M$, for any $\lambda \in \Omega_-$, we have

$$\begin{aligned}
&P\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}}) > 0\} \\
&\geq P\{\inf_{\lambda\in\Omega_-} \frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_\lambda})}{n} - \frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})}{n} - \frac{\kappa_n \mathrm{df}_{\bar{\alpha}}}{n} > 0\} \\
&= P\{\min_{\alpha\not\supseteq\alpha_0}[\frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha})}{n} - \log(n)\rho_1] - [\frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})}{n} - \log(n)\rho_1] - \frac{\kappa_n \mathrm{df}_{\bar{\alpha}}}{n} > 0\} \qquad \text{(S.5)} \\
&= P\{\min_{\alpha\not\supseteq\alpha_0} c_\alpha - c_{\bar{\alpha}} + o_P(1) > 0\} \to 1, \qquad \text{(S.6)}
\end{aligned}$$

as $n \to \infty$. (S.5) is due to the finiteness of $\mathcal{A}$, and (S.6) uses both (E4) and the fact that deviance tends to be smaller as covariate dimension increases. (S.1) follows from the above equations.

For any $\lambda \in \Omega_0, \alpha_\lambda = \alpha_0$, it follows by (2.5)

$$\begin{aligned}
&P\{\inf_{\lambda\in\Omega_0} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}}) > 0\} \\
&\geq P\{\inf_{\lambda\in\Omega_0} - 2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_\lambda}) - [-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})] - \kappa_n \mathrm{df}_{\bar{\alpha}} > 0\} \\
&= P\{-2[\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})] - \kappa_n \mathrm{df}_{\bar{\alpha}} > 0\} \\
&\geq P\{-2[\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})] > M \mathrm{df}_{\bar{\alpha}}\} \qquad \text{(S.7)} \\
&\to P\{\chi^2_{\mathrm{df}_{\bar{\alpha}} - \mathrm{df}_{\alpha_0}} \geq M \mathrm{df}_{\bar{\alpha}}\} > 0. \qquad \text{(S.8)}
\end{aligned}$$

(S.7) is due to $\kappa_n < M$, and (S.8) uses the fact that $\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}$ and $\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}}$ are asymptotically normal under regular condition (A)-(D) in Fan and Li (2002). Hence, the likelihood ratio test statistics $-2[\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})] \xrightarrow{\mathcal{L}} \chi^2_{\mathrm{df}_{\bar{\alpha}} - \mathrm{df}_{\alpha_0}}$. (S.2) follows by taking $\pi = P\{\chi^2_{\mathrm{df}_{\bar{\alpha}} - \mathrm{df}_{\alpha_0}} \geq$

$M\mathrm{df}_{\bar{\alpha}}\}$. This completes the proof of Lemma 1.

Lemma 2. *Suppose that the partial likelihood function of the Cox model satisfies Conditions (A)-(D) in Fan and Li (2002). Then under Condition (E1)-(E4), and let* $\lambda_n = \kappa_n/\sqrt{n}$. *If* κ_n *satisfies* $\kappa_n \to \infty$ *and* $\lambda_n \to 0$ *as* $n \to \infty$, *we have*

$$P\{GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) = GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha 0})\} \to 1, \tag{S.9}$$

$$P\left\{\inf_{\lambda \in (\Omega_- \cup \Omega_+)} GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) > GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n})\right\} \to 1. \tag{S.10}$$

Proof. With loss of generality, assume that the first d_{α_0} component of $\boldsymbol{\beta}_0$ are nonzero for the true model while the rest are zeros. By Conditions (A)-(D) in Fan and Li (2002) together with Condition (E3), Fan and Li (2002) showed that

$$\begin{gathered}\widehat{\beta}_{\lambda_n j} \xrightarrow{p} 0 \text{ for } j = d_{\alpha_0}+1, \cdots, d,\\ \tfrac{\partial}{\partial \beta_j}\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n j}) - p'_{\lambda_n}(|\widehat{\beta}_{\lambda_n j}|)\mathrm{sgn}(\widehat{\beta}_{\lambda_n j}) \xrightarrow{p} 0 \text{ for } j = 1, \cdots, d_{\alpha_0},\end{gathered} \tag{S.11}$$

where $\widehat{\beta}_{\lambda_n j}$ is the jth component of $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. Under Condition (E1) and (E2), for $j = 1, \cdots, d_{\alpha_0}$, there exits an m such that

$$p'_{\lambda_n}(|\widehat{\beta}_{\lambda_n j}|) = 0 \text{ for } |\widehat{\beta}_{\lambda_n j}| \geq \min\{|\beta_{\lambda_n j}|\} \geq m\lambda_n.$$

By (S.11), with probability tending to 1, we have,

$$\frac{\partial}{\partial \beta_j}\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n j}) = 0, \text{ for } j = 1, \cdots, d_{\alpha_0},$$

This is the score equation for the unpenalized partial likelihood under the true model α_0.

Therefore, with probability tending to 1, we have

$$\widehat{\boldsymbol{\beta}}_{\lambda_n} = \widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0},$$
$$\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n}) = \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}).$$

Thus $\mathrm{df}_{\alpha_{\lambda_n}} = \mathrm{df}_{\alpha_0}$ with probability tending to 1. Hence it follows that,

$$\begin{aligned}
&P\{\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) = \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0})\} \\
&= P\{-2\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n}) + \kappa_n \mathrm{df}_{\alpha_{\lambda_n}} + 2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}) - \kappa_n \mathrm{df}_{\alpha_0} = 0\} \\
&= P\{-2[\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0})] + \kappa_n(\mathrm{df}_{\alpha_{\lambda_n}} - \mathrm{df}_{\alpha_0}) = 0\} \\
&\to 1.
\end{aligned}$$

This validates (S.9).

Next, we want to show that $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) > \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n})$ for any λ that cannot result in the true model. First, we consider λ that could result in underfitting models, namely, $\lambda \in \Omega_- = \{\lambda : \alpha_\lambda \not\supseteq \alpha_0\}$. By (S.4) and (S.9), with probability tending to 1, it follows that

$$\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) > -2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_\lambda}) - [-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0})] - \kappa_n \mathrm{df}_{\alpha_0}.$$

For any $\lambda \in \Omega_- = \{\lambda : \alpha \not\supseteq \alpha_0\}$, we can take $\inf_{\lambda \in \Omega_-}$ over $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda})$. Under Condition (E4) and $\kappa_n/\sqrt{n} \to 0$, for any $\lambda \in \Omega_-$, we have

$$
\begin{aligned}
& P\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) > 0\} \\
& \geq P\{\inf_{\lambda\in\Omega_-} \frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda})}{n} - \frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} - \frac{\kappa_n \mathrm{df}_{\alpha_0}}{n} > 0\} \\
& = P\{\min_{\alpha\not\supseteq\alpha_0}[\frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha})}{n} - \log(n)\rho_1] - [\frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} - \log(n)\rho_1] - \frac{\kappa_n \mathrm{df}_{\alpha_0}}{n} > 0\} \\
& = P\left\{\min_{\alpha\not\supseteq\alpha_0} c_\alpha - c_{\alpha_0} + o_P(1) > 0\right\} \to 1, \qquad \text{(S.12)}
\end{aligned}
$$

as $n \to \infty$. (S.12) is due to Condition (E4). This implies that

$$
P\left\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n})\right\} \to 1. \qquad \text{(S.13)}
$$

For any $\lambda \in \Omega_+ = \{\lambda : \alpha_\lambda \supset \alpha_0\}$, we have

$$
\begin{aligned}
& \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) \\
& = -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) - [-2\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n})] + \kappa_n(\mathrm{df}_{\alpha_\lambda} - \mathrm{df}_{\alpha_{\lambda_n}}) \\
& \geq -2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}) - [-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_0})] + \kappa_n\tau_n, \qquad \text{(S.14)}
\end{aligned}
$$

where $\tau_n > 0$ due to the fact that $\mathrm{df}_{\alpha_\lambda} - \mathrm{df}_{\alpha_{\lambda_n}} = \tau_n > 0$ when n is large. And (S.14) follows (S.3). We then take $\inf_{\lambda\in\Omega_+}$ over $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda)$. Under Condition (E4) and $\kappa_n/\sqrt{n} \to 0$, for any $\lambda \in \Omega_+$, we have

$$
\begin{aligned}
& \inf_{\lambda\in\Omega_+} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) \\
& \geq \min_{\alpha\not\supseteq\alpha_0} -2[\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha}) - \ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_0})] + \kappa_n\tau_n \qquad \text{(S.15)} \\
& = \kappa_n\tau_n\{1 + o_p(1)\}. \qquad \text{(S.16)}
\end{aligned}
$$

(S.16) uses the fact that $2[\ell_c(\widehat{\boldsymbol{\beta}}^\star_\alpha) - \ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_0})] \to \chi^2_{\mathrm{df}_\alpha - \mathrm{df}_{\alpha_0}}$ for $\alpha \supset \alpha_0$ together with that $\kappa_n \to \infty$. Therefore, (S.15) is positive as $n \to \infty$. Hence, we have,

$$P\left\{\inf_{\lambda\in\Omega_+} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n})\right\} \to 1. \tag{S.17}$$

Based on (S.13) and (S.17) together, we prove (S.10). Consequently, this completes the proof of Lemma 2.

Proofs of Theorem 2. Lemma 1 implies that for any λ producing the underfitted model, its associated $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda)$ is consistently larger than $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^\star_{\bar{\alpha}})$. Thus, the optimal model selected by minimizing the $\mathrm{GIC}_{\kappa_n}(\boldsymbol{\beta})$ must be either the true model or overfitted models with probability tending to one. In addition, Lemma 1 indicates that there is a nonzero probability that the smallest value of $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda)$ associated with the true model is larger than that of the full model. As a result, there is a positive probability that any λ associated with the true model cannot be selected by $\mathrm{GIC}_{\kappa_n}(\boldsymbol{\beta})$ as the regularization parameter. Theorem 2(A) follows.

Lemma 2 indicates that the model identified by λ_n converges to the true model as the sample size gets large. In addition, it shows that those λ's, which fail to identify the true model, cannot be selected by $\mathrm{GIC}_{\kappa_n}(\boldsymbol{\beta})$ asymptotically. Theorem 2(B) follows.

We next show Theorem 2(C). Note that $(1 - \mathrm{df}_\lambda/n)^2 = 1 + 2\mathrm{df}_\lambda/n + O(\{\mathrm{df}_\lambda/n\}^2)$. By the definition of the GCV, it follows that

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

Theorem 1 implies $-\ell_c(\widehat{\boldsymbol{\beta}}_\lambda)/(n\log(n)) \to \rho_1 > 0$ as $n \to \infty$, then

$$\begin{aligned} 2n\mathrm{GCV}(\lambda) &= -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + 4\rho_1\log(n)\mathrm{df}_\lambda\{1 + o_p(1)\} + o_p(1) \\ &= -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + \kappa_{gcv}\mathrm{df}_{\alpha_\lambda}\{1 + o_p(1)\}, \end{aligned}$$

where $\kappa_{gcv} = 4\rho_1 \log(n)$. Theorem 2(C) follows by using the following the proof of Lemma 2.