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

\[
P\left\{ \inf_{\lambda \in \Omega} GIC_{\kappa_n}(\hat{\beta}_\lambda) > GIC_{\kappa_n}(\hat{\beta}^*_\lambda) \right\} \to 1 \quad \text{as} \quad n \to \infty, \quad (S.1)
\]

\[
\liminf_{n \to \infty} P\left\{ \inf_{\lambda \in \Omega_0} GIC_{\kappa_n}(\hat{\beta}_\lambda) > GIC_{\kappa_n}(\hat{\beta}^*_\lambda) \right\} \geq \pi. \quad (S.2)
\]

**Proof.** Recall that for any given \(\lambda\), we can obtain a selected model \(\alpha_\lambda\) by penalized variable selection. And based on this selected model \(\alpha_\lambda\), we are able to obtain its corresponding non-penalized estimates \(\hat{\beta}^*_\alpha_\lambda\) by maximizing the corresponding partial likelihood. Then

\[
\ell_c(\hat{\beta}^*_\alpha_\lambda) \geq \ell_c(\hat{\beta}_\lambda), \quad (S.3)
\]

and \(-2\ell_c(\hat{\beta}_\lambda) + \kappa_n df_\lambda > -2\ell_c(\hat{\beta}^*_\alpha_\lambda)\) Thus,

\[
GIC_{\kappa_n}(\hat{\beta}_\lambda) > -2\ell_c(\hat{\beta}^*_\alpha_\lambda). \quad (S.4)
\]
Subtract $\text{GIC}_{\kappa_{\alpha}}(\hat{\beta}_{\alpha}^*)$ from both size of (S.4), we can obtain that

$$\text{GIC}_{\kappa_{\alpha}}(\hat{\beta}_{\alpha}) - \text{GIC}_{\kappa_{\alpha}}(\hat{\beta}_{\alpha}^*) > -2\ell_{c}(\hat{\beta}_{\alpha}^*) - \{ -2\ell_{c}(\hat{\beta}_{\alpha}^*) + \kappa_{n}df_{\alpha} \}.$$

For any $\lambda \in \Omega_{-} = \{ \lambda : \alpha \not\subseteq \alpha_{0} \}$, we can take $\inf_{\lambda \in \Omega_{-}}$ over $\text{GIC}_{\kappa_{\alpha}}(\hat{\beta}_{\alpha})$. Under Condition (E4) and $\kappa_{n} < M$, for any $\lambda \in \Omega_{-}$, we have

$$P\{ \inf_{\lambda \in \Omega_{-}} \text{GIC}_{\kappa_{\alpha}}(\hat{\beta}_{\alpha}) - \text{GIC}_{\kappa_{\alpha}}(\hat{\beta}_{\alpha}^*) > 0 \}$$

$$\geq P\{ \inf_{\lambda \in \Omega_{-}} -\frac{2\ell_{c}(\hat{\beta}_{\alpha}^*)}{n} - \frac{-2\ell_{c}(\hat{\beta}_{\alpha}^*)}{n} - \frac{\kappa_{n}df_{\alpha}}{n} > 0 \}$$

$$= P\{ \min_{\alpha \subsetneq \alpha_{0}} \left[ -\frac{2\ell_{c}(\hat{\beta}_{\alpha}^*)}{n} - \log(n)\rho_{1} \right] - \left[ -\frac{2\ell_{c}(\hat{\beta}_{\alpha}^*)}{n} - \log(n)\rho_{1} \right] - \frac{\kappa_{n}df_{\alpha}}{n} > 0 \}$$

$$= P\{ \min_{\alpha \subsetneq \alpha_{0}} c_{\alpha} - c_{\alpha} + o_{P}(1) > 0 \} \to 1,$$  \hspace{1cm} (S.5)

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)

$$P\{ \inf_{\lambda \in \Omega_{0}} \text{GIC}_{\kappa_{\alpha}}(\hat{\beta}_{\alpha}) - \text{GIC}_{\kappa_{\alpha}}(\hat{\beta}_{\alpha}^*) > 0 \}$$

$$\geq P\{ \inf_{\lambda \in \Omega_{0}} -2\ell_{c}(\hat{\beta}_{\alpha_{0}}) - \{ -2\ell_{c}(\hat{\beta}_{\alpha}^*) \} - \kappa_{n}df_{\alpha} > 0 \}$$

$$= P\{ -2[\ell_{c}(\hat{\beta}_{\alpha_{0}}) - \ell_{c}(\hat{\beta}_{\alpha}^*)] - \kappa_{n}df_{\alpha} > 0 \}$$

$$\geq P\{ -2[\ell_{c}(\hat{\beta}_{\alpha_{0}}) - \ell_{c}(\hat{\beta}_{\alpha}^*)] > Mdf_{\alpha} \}$$

$$\to P\{ \chi_{df_{\alpha} - df_{\alpha_{0}}}^{2} \geq Mdf_{\alpha} \} > 0.$$  \hspace{1cm} (S.7)

(S.7) is due to $\kappa_{n} < M$, and (S.8) uses the fact that $\hat{\beta}_{\alpha_{0}}^*$ and $\hat{\beta}_{\alpha}^*$ are asymptotically normal under regular condition (A)-(D) in Fan and Li (2002). Hence, the likelihood ratio test statistics $-2[\ell_{c}(\hat{\beta}_{\alpha_{0}}) - \ell_{c}(\hat{\beta}_{\alpha}^*)] \xrightarrow{L} \chi_{df_{\alpha} - df_{\alpha_{0}}}^{2}$. (S.2) follows by taking $\pi = P\{ \chi_{df_{\alpha} - df_{\alpha_{0}}}^{2} \geq}$
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 \( \kappa_n \) satisfies \( \kappa_n \to \infty \) and \( \lambda_n \to 0 \) as \( n \to \infty \), we have

\[
P\{ \text{GIC}_{\kappa_n}(\hat{\beta}_n) = \text{GIC}_{\kappa_n}(\hat{\beta}_{a_0}) \} \to 1, \quad (S.9)
\]

\[
P \left\{ \inf_{\lambda \in (\Omega^- \cup \Omega^+)} \text{GIC}_{\kappa_n}(\hat{\beta}_n) > \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda_n}) \right\} \to 1. \quad (S.10)
\]

Proof. With loss of generality, assume that the first \( d_{a_0} \) component of \( \beta_0 \) 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

\[
\frac{\partial}{\partial \beta_j} \ell_c(\hat{\beta}_{\lambda_n}) - p_{\lambda_n}'(|\hat{\beta}_{\lambda_n}|) \text{sgn}(\hat{\beta}_{\lambda_n}) \overset{p}{\to} 0 \quad \text{for} \ j = 1, \cdots, d_{a_0},
\]

(S.11)

where \( \hat{\beta}_{\lambda_n} \) is the \( j \)th component of \( \hat{\beta}_n \). Under Condition (E1) and (E2), for \( j = 1, \cdots, d_{a_0} \), there exists an \( m \) such that

\[
p_{\lambda_n}'(|\hat{\beta}_{\lambda_n}|) = 0 \quad \text{for} \ |\hat{\beta}_{\lambda_n}| \geq \min\{|\hat{\beta}_{\lambda_n}|\} \geq m \lambda_n.
\]

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

\[
\frac{\partial}{\partial \beta_j} \ell_c(\hat{\beta}_{\lambda_n}) = 0, \quad \text{for} \ j = 1, \cdots, d_{a_0},
\]

This is the score equation for the unpenalized partial likelihood under the true model \( a_0 \).
Therefore, with probability tending to 1, we have

\[ \hat{\beta}_{\lambda_n} = \hat{\beta}_{\alpha_0}^*, \]
\[ \ell_c(\hat{\beta}_{\lambda_n}) = \ell_c(\hat{\beta}_{\alpha_0}^*). \]

Thus df_{\alpha_{\lambda_n}} = df_{\alpha_0} with probability tending to 1. Hence it follows that,

\[ P\{ \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda_n}) = \text{GIC}_{\kappa_n}(\hat{\beta}_{\alpha_0}^*) \} \]
\[ = P\{ -2\ell_c(\hat{\beta}_{\lambda_n}) + \kappa_n df_{\alpha_{\lambda_n}} + 2\ell_c(\hat{\beta}_{\alpha_0}^*) - \kappa_n df_{\alpha_0} = 0 \} \]
\[ = P\{ -2[\ell_c(\hat{\beta}_{\lambda_n}) - \ell_c(\hat{\beta}_{\alpha_0}^*)] + \kappa_n(df_{\alpha_{\lambda_n}} - df_{\alpha_0}) = 0 \} \]
\[ \to 1. \]

This validates (S.9).

Next, we want to show that \( \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda}) > \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda_n}) \) for any \( \lambda \) that cannot result in the true model. First, we consider \( \lambda \) 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

\[ \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda}) - \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda_n}) > -2\ell_c(\hat{\beta}_{\alpha_{\lambda}}^*) - [2\ell_c(\hat{\beta}_{\alpha_0}^*)] - \kappa_n df_{\alpha_0}. \]

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

\[
P\{ \inf_{\lambda \in \Omega_{-}} \text{GIC}_{\kappa_n}(\hat{\beta}_\lambda) - \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda_n}) > 0 \}
\geq P\{ \inf_{\lambda \in \Omega_{-}} \frac{-2\ell_c(\hat{\beta}_{\alpha_\lambda})}{n} - \frac{-2\ell_c(\hat{\beta}_{\alpha_0})}{n} - \frac{\kappa_n \text{df}_{\alpha_0}}{n} > 0 \}
= P\{ \min_{\alpha \neq \alpha_0} \left[ \frac{-2\ell_c(\hat{\beta}_\alpha)}{n} - \log(n)\rho_1 \right] - \left[ \frac{-2\ell_c(\hat{\beta}_{\alpha_0})}{n} - \log(n)\rho_1 \right] - \frac{\kappa_n \text{df}_{\alpha_0}}{n} > 0 \}
= P\left\{ \min_{\alpha \neq \alpha_0} c_{\alpha} - c_{\alpha_0} + o_P(1) > 0 \right\} \to 1,
\]  
(S.12)

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

\[
P\left\{ \inf_{\lambda \in \Omega_{-}} \text{GIC}_{\kappa_n}(\hat{\beta}_\lambda) > \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda_n}) \right\} \to 1.
\]  
(S.13)

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

\[
\text{GIC}_{\kappa_n}(\hat{\beta}_\lambda) - \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda_n})
= -2\ell_c(\hat{\beta}_\lambda) - \left[ -2\ell_c(\hat{\beta}_{\lambda_n}) \right] + \kappa_n (\text{df}_{\alpha_\lambda} - \text{df}_{\alpha_{\lambda_n}})
\geq -2\ell_c(\hat{\beta}_{\alpha_\lambda}) - \left[ -2\ell_c(\hat{\beta}_{\alpha_0}) \right] + \kappa_n \tau_n,
\]  
(S.14)

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

\[
\inf_{\lambda \in \Omega_{+}} \text{GIC}_{\kappa_n}(\hat{\beta}_\lambda) - \text{GIC}_{\kappa_n}(\hat{\beta}_{\lambda_n})
\geq \min_{\alpha \neq \alpha_0} -2[\ell_c(\hat{\beta}_\alpha) - \ell_c(\hat{\beta}_{\alpha_0})] + \kappa_n \tau_n
\geq \kappa_n \tau_n \{ 1 + o_P(1) \}.
\]  
(S.15)
(S.16) uses the fact that \(2[\ell_c(\widehat{\beta}^\star) - \ell_c(\widehat{\beta}_{\alpha_0}^\star)] \rightarrow \chi^2_{df_{\alpha} - df_{\alpha_0}}\) for \(\alpha \supset \alpha_0\) together with that \(\kappa_n \rightarrow \infty\). Therefore, (S.15) is positive as \(n \rightarrow \infty\). Hence, we have,

\[
P\left\{ \inf_{\lambda \in \Omega^+} \text{GIC}_{\kappa_n}(\widehat{\beta}_\lambda) > \text{GIC}_{\kappa_n}(\widehat{\beta}_{\lambda_n}) \right\} \rightarrow 1. \tag{S.17}
\]

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 \(\lambda\) producing the underfitted model, its associated \(\text{GIC}_{\kappa_n}(\widehat{\beta}_\lambda)\) is consistently larger than \(\text{GIC}_{\kappa_n}(\widehat{\beta}_{\star})\). Thus, the optimal model selected by minimizing the \(\text{GIC}_{\kappa_n}(\beta)\) 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 \(\text{GIC}_{\kappa_n}(\widehat{\beta}_\lambda)\) associated with the true model is larger than that of the full model. As a result, there is a positive probability that any \(\lambda\) associated with the true model cannot be selected by \(\text{GIC}_{\kappa_n}(\beta)\) as the regularization parameter. Theorem 2(A) follows.

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

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

\[
2n\text{GCV}(\lambda) = -2\ell_c(\widehat{\beta}_\lambda) + 4(-\ell_c(\widehat{\beta}_\lambda)/n)df_{\lambda} + O_p\left\{df_{\lambda}/n\right\}^2\ell_c(\widehat{\beta}_\lambda)
\]

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

\[
2n\text{GCV}(\lambda) = -2\ell_c(\widehat{\beta}_\lambda) + 4\rho_1 \log(n)df_{\lambda} \{1 + o_p(1)\} + o_p(1)
\]

\[= -2\ell_c(\widehat{\beta}_\lambda) + \kappa_{gcv} df_{\lambda} \{1 + o_p(1)\},\]
where $\kappa_{gcv} = 4 \rho_1 \log(n)$. Theorem 2(C) follows by using the following the proof of Lemma 2.