SIMULTANEOUS VARIABLE SELECTION AND ESTIMATION OF SURVIVAL MODEL WITH INFORMATIVE CENSORING

Zili Liu[1], Hong Wang[1], Chunjie Wang[2] and Xinyuan Song[3*]

[1] Central South University

[2] Changchun University of Technology

[3] The Chinese University of Hong Kong

Supplementary Material

In the supplementary material, we will sketch the proof of the asymptotic properties described in Theorems 1 to 3, Algorithm 1, an extended BIC and additional numerical results in Tables S1-S6.

S1 Proofs of Theorems

S1.1 Notations

Without loss of generality, we assume regression coefficients in $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are all bounded by a constant C. We denote the space for $\boldsymbol{\beta}$ by $B_{\boldsymbol{\beta}} = \{\boldsymbol{\beta} : |\beta_j| \leq C < \infty, \forall j\}$, and the space for $\boldsymbol{\phi}$ by $B_{\boldsymbol{\phi}} = \{\boldsymbol{\phi} : |\phi_j| \leq C < \infty, \forall j\}$. To simplify notations we let $a = X_{(1)}$ and $b = X_{(n)}$. Assume that functions $h_{0T}(x)$ and $h_{0C}(x)$ are bounded and have $r(\geq 1)$ continuous deriva-

tives over [a, b]. Let $C^r[a,b]$ denote the set for these functions. The spaces for $h_{0T}(x)$ and $h_{0C}(x)$ are $A_T = \{h_{0T}(x) : h_{0T}(x) \in C^r[a,b], 0 \leq h_{0T}(x) \leq C_1 < \infty, \forall x \in [a,b]\}$ and $A_C = \{h_{0C}(x) : h_{0C}(x) \in C^r[a,b], 0 \leq h_{0C}(x) \leq C_1 < \infty, \forall x \in [a,b]\}$ respectively. Then the parameter space for $(\boldsymbol{\beta}, \boldsymbol{\phi}, h_{0T}(x), h_{0C}(x))$ is $\Pi = B_{\boldsymbol{\beta}} * B_{\phi} * A_T * A_C$. In this section, we denote the piece-wise constant approximating functions to $h_{0T}(x)$ and $h_{0C}(x)$ by $h_{nT}(x)$ and $h_{nC}(x)$ respectively, and they are written as $h_{nT}(x) = \sum_{u=1}^{m} \theta_{nu} I\left(x \in \mathcal{B}_u\right)$ and $h_{nC}(x) = \sum_{u=1}^{m} \gamma_{nu} I\left(x \in \mathcal{B}_u\right)$, where θ_{nu} and γ_{nu} are assumed bounded and non-negative. $A_{nT} = \{0 \leq h_{nT}(x) \leq C_2 < \infty, \forall x \in [a,b]\}$ and $A_{nC} = \{0 \leq h_{nC}(x) \leq C_2 < \infty, \forall x \in [a,b]\}$ are the spaces for $h_{nT}(x)$ and $h_{nC}(x)$. The parameter space for $(\boldsymbol{\beta}, \boldsymbol{\phi}, h_{nT}(x), h_{nC}(x))$ is $\Pi_n = B_{\boldsymbol{\beta}} * B_{\boldsymbol{\phi}} * A_{nT} * A_{nC}$. Let $\boldsymbol{\pi} = (\boldsymbol{\beta}, \boldsymbol{\phi}, h_{0T}(x), h_{0C}(x))$ and $\boldsymbol{\pi}_n = (\boldsymbol{\beta}, \boldsymbol{\phi}, h_{nT}(x), h_{nC}(x))$. The MPL estimators are denoted by $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{h}_{nT}(x)$ and $\hat{h}_{nC}(x)$, where $\hat{h}_{nT}(x) = \sum_{u=1}^{m} \hat{\theta}_{nu} I\left(x \in \mathcal{B}_u\right)$ and $\hat{h}_{nC}(x) = \sum_{u=1}^{m} \hat{\gamma}_{nu} I(x \in \mathcal{B}_u)$.

Let $\boldsymbol{W}_i = \left(X_i, \delta_i, \boldsymbol{Z}_i^{\mathrm{T}}\right)$ for $i = 1, \cdots, n$, and we assumed they are i.i.d. Let $\boldsymbol{W}$ be a general $\boldsymbol{W}_i$ and $F(\boldsymbol{w}; \boldsymbol{\pi})$ be the cumulative distribution function of $\boldsymbol{W}$. Corresponding to the parameter spaces Π and Π_n, the log-likelihood functions are denoted by $\ell(\boldsymbol{\pi}; \boldsymbol{W})$ and $\ell\left(\boldsymbol{\pi}_n; \boldsymbol{W}\right)$ respectively. Let $\boldsymbol{\pi}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, h_{0T}^0(x), h_{0C}^0(x))$. We define $P\ell(\boldsymbol{\pi}) = \int \ell(\boldsymbol{\pi}; \boldsymbol{w}) dF\left(\boldsymbol{w}; \boldsymbol{\pi}_0\right) \equiv E_0(\ell(\boldsymbol{\pi}; \boldsymbol{W}))$ for $\boldsymbol{\pi} \in \Pi$ and $P_n\ell(\boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(\boldsymbol{\pi}; \boldsymbol{W}_i\right)$. For $\boldsymbol{\pi}_n \in \Pi_n . P\ell\left(\boldsymbol{\pi}_n\right)$

and $P_n\ell\left(\boldsymbol{\pi}_n\right)$ are similarly defined. Clearly $\boldsymbol{\pi}_0$ maximizes $P\ell(\boldsymbol{\pi})$. If $\ell(\boldsymbol{\pi};\boldsymbol{W})$ is the true log-likelihood (therefore, the copula function correctly captures the dependence between T and C), then $\boldsymbol{\pi}_0$ is the true $\boldsymbol{\pi}$ that generates the data.

S1.2 Assumptions

We first give some assumptions required for consistent results.

A1. Matrix $\boldsymbol{Z}$ is bounded and $E\left(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right)$ is non-singular.

A2. The penalty function $J(\cdot)$ is bounded.

A3. For functions $h_{nT}(x)$ and $h_{nC}(x)$, their corresponding coefficient vectors $\boldsymbol{\theta}_n$ and $\boldsymbol{\gamma}_n$ are in a compact subset of R^m, and their elements are bounded.

A4. Assume for any $h_{0T}(x)$ and $h_{0C}(x)$ in $C^r[a,b]$, we let $a = X_{(1)}$ and $b = X_{(n)}$ there exist $h_{nT}(x)$ and $h_{nC}(x)$ such that $\max_x \mid h_{nT}(x) - h_{0T}(x) \mid \to 0$ and $\max_x |h_{nC}(x) - h_{0C}(x)| \to 0$ as $m \to \infty$ and $n \to \infty$ but $m/n \to 0$.

A5. For k = 1, 2, denote $h_{n1}(\cdot) = h_{0T}(\cdot)$ and $h_{n2}(\cdot) = h_{0C}(\cdot)$. Suppose the $h_{nk}(\cdot)$ is Lip-α continuous with exponent M. That is, there exist a positive real value M and some $\alpha \in (0,1]$ such that $|h_{nk}(t_1) - h_{nk}(t_2)| \leq$

$M|t_1 - t_2|^\alpha$ for all $t_1, t_2 \in (l, u), k = 1, 2$.

A6. For $k = 1, 2$, $\iota_k\sqrt{n^{1-\nu}/p} \to \infty$ with $0 < \nu < 1/2$, where $\iota_k = \min_{j \in \mathscr{C}_k} |\eta_{10,j}|$.

The following assumptions are needed to show that the estimator is asymptotically normal and efficient:

B.1 Random vectors $\left(X_i, \delta_i, \boldsymbol{Z}_i^{\mathrm{T}}\right), 1 \le i \le n$, are independent and identically distributed, and the distribution of $\boldsymbol{Z}_i$ is independent of $\boldsymbol{\eta}$.

B.2 Let Ω be the parameter space for $\boldsymbol{\eta}$ where Ω is a compact subset of $R^{2(p+m)}$. Assume $E_{\eta_0}\left(n^{-1}\ell(\boldsymbol{\eta})\right)$ exists and has a unique maximum at $\boldsymbol{\eta}_0 \in \Omega$.

B.3 Both $J_1(\boldsymbol{\eta}) = J(\boldsymbol{\theta})$ and $J_2(\boldsymbol{\eta}) = J(\boldsymbol{\gamma})$ are continuous and bounded over Ω, and their first two derivatives exist for all $\boldsymbol{\eta} \in \Omega$. Moreover, their second derivatives are bounded in a neighborhood of $\boldsymbol{\eta}_0$.

B.4 Assume $\ell(\boldsymbol{\eta})$ is bounded and is twice continuously differentiable in a neighborhood of $\boldsymbol{\eta}_0$, and the matrices

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}} \text{ and } \lim_{n\to\infty} n^{-1} \frac{\partial^2 \ell(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^{\mathrm{T}}}$$

exists.

B.5 Assume there are q active constraints from $\boldsymbol{\theta}$ and l actives constraints from $\boldsymbol{\gamma}$. Lelt $\boldsymbol{G}_0(\boldsymbol{\eta}) = -n^{-1}E_{\boldsymbol{\eta}_0}\partial^2\ell(\boldsymbol{\eta})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}} + \mu_{1n}\partial^2 J_1(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}} + \mu_{2n}\partial^2 J_2(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}} + \lambda\partial^2 p_\lambda(\boldsymbol{\eta_1})/\partial\boldsymbol{\eta_1}\partial\boldsymbol{\eta_1}^{\mathrm{T}}$. Let $\boldsymbol{U}_{[2(m+p)]\times[s_1+s_2+2p-q-l]}$ be a matrix whose rows corresponding to active constraints take zero values, and other rows form an identity matrix. Assume $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{G}_0(\boldsymbol{\eta})\boldsymbol{U}$ is invertible in a neighborhood of $\boldsymbol{\eta}_0$.

S1.3 Preliminaries

Assuming that for any $\boldsymbol{\pi} \in \Pi$, there exists an $\boldsymbol{\pi}_n \in \Pi_n$ such that $\rho(\boldsymbol{\pi}_n, \boldsymbol{\pi}) \to 0$ when $n \to \infty$, where

$$\begin{aligned}\rho(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) =& \{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2\|_2^2 + \sup_{x\in[a,b]} \left|h_{0T}^1(x) - h_{0T}^2(x)\right|^2 \\ &+ \sup_{x\in[a,b]} \left|h_{0C}^1(x) - h_{0C}^2(x)\right|^2\}^{1/2}.\end{aligned} \tag{A.1}$$

This assumption can be guaranteed if the bins are placed to cover the entire interval $[a, b]$, and the number of bins has to grow to ∞ slower than $n \to \infty$.

We wish to adopt Xue, Lam and Li (2004) to develop strong consistency properties for the MPL estimate $\hat{\boldsymbol{\pi}}_n$. Their derivations are developed based on the concept of sieve maximum likelihood estimate (sieve-MLE). In fact, it is easy to explain that our MPL estimate is also a sieve-MLE under certain conditions. Let $\lambda_n = \lambda_0/n$, $\mu_{1n} = h_1/n, \mu_{2n} = h_2/n$ and $\hat{\boldsymbol{\pi}}_n$

be the MPL estimate of $\boldsymbol{\pi}_n \in \Pi_n$ maximizing $P_n\ell(\boldsymbol{\pi}_n) - \lambda_n p_{mic}(\boldsymbol{\beta}, \boldsymbol{\phi}) - \mu_{1n}J(h_{nT}) - \mu_{2n}J(h_{nC})$. Since $P_n\ell(\hat{\boldsymbol{\pi}}_n) \geq P_n\ell(\boldsymbol{\pi}_n^*) - \varepsilon_n$, where $\boldsymbol{\pi}_n^* = (\boldsymbol{\beta}^*, \boldsymbol{\phi}^*, h_{0T}^*(x), h_{0C}^*(x))$ and $\boldsymbol{\eta}_1 = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\phi}^{\mathrm{T}})^{\mathrm{T}}$ represents the MLE of $\boldsymbol{\pi}_n$ and

$$\begin{aligned}\varepsilon_n = \lambda_n| - p_{mic}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}) + p_{mic}(\boldsymbol{\beta}^*, \boldsymbol{\phi}^*)| + \mu_{1n}| \\ - J(\hat{h}_{nT}) + J(h_{nT}^*)| + \mu_{2n}| - J(\hat{h}_{nC}) + J(h_{nC}^*)| \to 0\end{aligned}$$

when both λ_n, μ_{1n} and μ_{2n} converges to 0 (note that J is bounded under Assumption A4), $\hat{\boldsymbol{\pi}}_n$ is a sieve-MLE according to the definition. The following results consider consistency of the MPL estimates of $\boldsymbol{\beta}, \boldsymbol{\phi}, h_{nT}(x)$ and $h_{nC}(x)$.

S1.4 Proofs

Proof of Theorem 1. According to the formula (A.1), the result of Theorem 1 can be demonstrated by showing that $\rho(\hat{\boldsymbol{\pi}}_n, \boldsymbol{\pi}_0) \to 0$ (a.s.) as $n \to \infty$, where $\hat{\boldsymbol{\pi}}_n$ and $\boldsymbol{\pi}_0$ have been defined in Section 3. Since λ_n, μ_{1n} and $\mu_{2n} \to 0$ when $n \to \infty$ and the penalty functions are bounded, the proof for consistency can be concentrated on the log-likelihood function only.

Firstly, we show that $N(\epsilon, A_{nT} * A_{nC}, L_\infty * L_\infty) \leq (12C_3C_4/\epsilon)^{2m}$ where C_3 and C_4 are constants specified below. For any $h_{nT1}, h_{nT2} \in A_{nT}$ and any $h_{nC1}, h_{nC2} \in A_{nC}$, where $h_{nTj}(x) = \sum_u \theta_{un}^j I(x \in \mathcal{B}_u)$ and $h_{nCj}(x) =$

$\sum_u \gamma_{un}^j I\left(x \in \mathcal{B}_u\right)$ with $j = 1, 2$, we have

$$\begin{aligned}
&\max_x |h_{nT1}(x) - h_{nT2}(x)| + \max_x |h_{nC1}(x) - h_{nC2}(x)| \\
&\quad \leq C_3(\max_u |\theta_{un}^1 - \theta_{un}^2| + \max_u |\gamma_{un}^1 - \gamma_{un}^2|) \\
&\quad \leq C_3(\|\theta^1 - \theta^2\|_2 + \|\gamma^1 - \gamma^2\|_2),
\end{aligned}$$

where $C_3 > 0$ is the upper bound of $\sum_u I\left(x \in \mathcal{B}_u\right)$ and θ^j and γ^j are m vectors with elements θ_{un}^j and γ_{un}^j respectively.

Then we can apply Lemma 4.1 of Pollard (1984) or Lemma A.2 of Xue, Lam and Li (2004) to give

$$N\left(\epsilon, \left\{0 \leq \pi_{un} \leq C_4, 1 \leq u \leq m\right\}, L_2\right) \leq \left(6C_4/\epsilon\right)^m.$$

Let $C_4 > 0$ be a common finite upper bound for both θ_{un} and γ_{un}. Then, we have

$$\begin{aligned}
&N\left(\epsilon, A_{nT} * A_{nC}, L_\infty * L_\infty\right) \\
&\leq N\left(\frac{\epsilon}{C_3}, \left\{0 \leq \theta_{un} \leq C_4, 1 \leq u \leq m\right\} * \left\{0 \leq \gamma_{un} \leq C_4, 1 \leq u \leq m\right\}, L_2 * L_2\right) \\
&\leq N\left(\frac{\epsilon}{2C_3}, \left\{0 \leq \theta_{un} \leq C_4, 1 \leq u \leq m\right\}, L_2\right) N\left(\frac{\epsilon}{2C_3}, \left\{0 \leq \gamma_{un} \leq C_4, 1 \leq u \leq m\right\}, L_2\right) \\
&\leq \left(12C_3C_4/\epsilon\right)^{2m}.
\end{aligned}$$

Let $\mathcal{L}_n = \{\ell(\boldsymbol{\pi}_n); \boldsymbol{\pi}_n \in \Pi_n\}$. In fact, similar to the proof of Lemma A.2 in

Xue, Lam and Li (2004), we have

$$N(\epsilon, \mathcal{L}_n, L_\infty) \leq N(\epsilon/3, B_{\boldsymbol{\beta}}, L_2) N(\epsilon/3, B_{\boldsymbol{\phi}}, L_2) N(\epsilon/3, A_{nT} * A_{nC}, L_\infty * L_\infty)$$
$$\leq (18C/\epsilon)^{2p} (36C_3C_4/\epsilon)^{2m} = K/\epsilon^{2(p+m)},$$

where $K = 2^{2p+4m} 3^{4p+4m} C^{2p} C_3^{2m} C_4^{2m}$. Next, define $\alpha_n = n^{-1/2+a}\sqrt{\log n}$ where $a \in (b/2, 1/2)$ with $b < 1$, and define $\epsilon_n = \epsilon\alpha_n$ for a fixed ϵ. From the proof of Theorem 1 of Xue, Lam and Li (2004), we can show that $\text{var}[P_n \ell(\boldsymbol{\pi}_n)] = o(8\epsilon_n^2)$ for any $\boldsymbol{\pi}_n \in \Pi_n$.

From Lemma 33 of Pollard (1984) we have, for a sufficient large n,

$$\begin{aligned}
&P\left(\sup_{\Pi_n} |P_n \ell(\boldsymbol{\pi}_n) - P\ell(\boldsymbol{\pi}_n)| > 8\epsilon_n\right) \\
&\leq 8N(\epsilon_n, \mathcal{L}_n, L_\infty) \exp\left(-\frac{1}{128} n\epsilon_n^2\right) \\
&\leq 8K \left(\frac{1}{\epsilon_n}\right)^{2(p+m)} \exp\left(-\frac{1}{128} n\epsilon^2\alpha_n^2\right) \\
&\leq 8K \exp\left\{2(p+m)\log \frac{1}{\epsilon n^{-1/2+a}(\log n)^{1/2}} - \frac{1}{128} n\epsilon^2 n^{-1+2a}\log n\right\} \\
&\leq 8K \exp\left\{-\left[\frac{1}{128}\epsilon^2 + (p+m)\frac{\log(\epsilon^2 n^{-1+2a}\log n)}{n^{2a}\log n}\right] n^{2a}\log n\right\} \\
&\leq 8Ke^{-K' n^{2a}\log n},
\end{aligned}$$

where K is the constant defined above and $K' > 0$ is a lower bound for the expression in the bracket of the exponential function. Existence of K' is guaranteed by the fact that the second term in the bracket converges to 0.

Therefore,

$$\sum_{n=1}^{\infty} P\left\{\sup_{\Pi_n} |P_n\ell\left(\boldsymbol{\pi}_n\right) - P\ell\left(\boldsymbol{\pi}_n\right)| > 8\epsilon_n\right\} < +\infty.$$

By the Borel-Cantelli lemma, we have

$$\sup_{\Pi_n} |P_n\ell\left(\boldsymbol{\pi}_n\right) - P\ell\left(\boldsymbol{\pi}_n\right)| \to 0 \quad \text{a.s. under measure } P. \tag{A.2}$$

Let $\boldsymbol{\pi}_{0n} = (\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, h_{0nT}(x), h_{0nC}(x)) \in \Pi_n$, which satisfies $\rho\left(\boldsymbol{\pi}_{0n}, \boldsymbol{\pi}_0\right) \to 0$ according to the Assumption A4, and $\rho\left(\cdot,\cdot\right)$ is defined in (A.1). Since $\boldsymbol{\pi}_0$ maximizes $P\ell(\boldsymbol{\pi})$ for $\boldsymbol{\pi} \in \Pi$ and $\hat{\boldsymbol{\pi}}_n$ maximizes $P_n\ell(\boldsymbol{\pi})$ for $\boldsymbol{\pi} \in \Pi_n$, we have

$$\begin{aligned}
&P_n\ell(\boldsymbol{\pi}_{0n}, \boldsymbol{W}) - P\ell(\boldsymbol{\pi}_{0n}, \boldsymbol{W}) + P\ell(\boldsymbol{\pi}_{0n}, \boldsymbol{W}) - P\ell(\boldsymbol{\pi}_0, \boldsymbol{W}) \\
&\quad \leq P_n\ell(\hat{\boldsymbol{\pi}}_n, \boldsymbol{W}) - P\ell(\boldsymbol{\pi}_0, \boldsymbol{W}) \\
&\quad \leq P_n\ell(\hat{\boldsymbol{\pi}}_n, \boldsymbol{W}) - P\ell(\hat{\boldsymbol{\pi}}_n, \boldsymbol{W}).
\end{aligned}$$

From the result of (A.2), we have $P_n\ell(\boldsymbol{\pi}_{0n}, \mathbf{W}) - P\ell(\boldsymbol{\pi}_{0n}, \mathbf{W}) \to 0$ and $P_n\ell(\hat{\boldsymbol{\pi}}_n, \mathbf{W}) - P\ell(\hat{\boldsymbol{\pi}}_n, \mathbf{W}) \to 0$ almost surely. In choosing the sieve space, we have $P\ell(\boldsymbol{\pi}_{0n}, \boldsymbol{W}) - P\ell(\boldsymbol{\pi}_0, \boldsymbol{W}) \to 0$, which can be established from $\rho(\boldsymbol{\pi}_{0n}, \boldsymbol{\pi}_0) \to 0$ and the fact that $\ell(\cdot)$ is continuous and bounded. Hence $|\, P_n\ell(\hat{\boldsymbol{\pi}}_n, \boldsymbol{W}) - P\ell(\boldsymbol{\pi}_0, \boldsymbol{W})\,| \to 0$ almost surely.

Moreover, We can show that $P\ell(\hat{\boldsymbol{\pi}}_n) - P\ell\left(\boldsymbol{\pi}_0\right)$ converges to 0 almost

surely from the fact

$$|P\ell(\hat{\boldsymbol{\pi}}_n, \boldsymbol{W}) - P\ell(\boldsymbol{\pi}_0, \boldsymbol{W})|$$
$$\leq |P\ell(\hat{\boldsymbol{\pi}}_n, \boldsymbol{W}) - P_n\ell(\hat{\boldsymbol{\pi}}_n, \boldsymbol{W})| + |P_n\ell(\hat{\boldsymbol{\pi}}_n, \boldsymbol{W}) - P\ell(\boldsymbol{\pi}_0, \boldsymbol{W})| \to 0 \quad \text{a.s.} \tag{A.3}$$

Let $q(\boldsymbol{w}; \boldsymbol{\pi})$ denote the Frchet derivative of the density function of $\boldsymbol{w}$ (i.e. $f(\boldsymbol{w}; \boldsymbol{\pi})$) with respect to $\boldsymbol{\pi}$. Let $\boldsymbol{\tau}$ be a point in between $\hat{\boldsymbol{\pi}}_n$ and $\boldsymbol{\pi}_0$. Hence $q(\boldsymbol{w}; \boldsymbol{\tau})$ is non-zero as $\boldsymbol{\tau}$ is not the MPL estimate. Note that both $q(\boldsymbol{w}; \boldsymbol{\pi})$ and $f(\boldsymbol{w}; \boldsymbol{\pi})$ are bounded. Because the Kullback-Leibler information is not less than the square of the Hellinger distance (Wong and Shen, 1995), we have

$$\begin{aligned}
&|P\ell(\hat{\boldsymbol{\pi}}_n; \boldsymbol{W}) - P\ell(\boldsymbol{\pi}_0; \boldsymbol{W})| = E(\ell(\boldsymbol{\pi}_0; \boldsymbol{W}) - \ell(\hat{\boldsymbol{\xi}}_n; \boldsymbol{W})) \\
&\geq \left\| f^{\frac{1}{2}}(\boldsymbol{W}; \boldsymbol{\pi}_0) - f^{\frac{1}{2}}(\boldsymbol{W}; \hat{\boldsymbol{\pi}}_n) \right\|_2^2 = \left\| \frac{q(\boldsymbol{W}; \boldsymbol{\tau})}{2f^{\frac{1}{2}}(\boldsymbol{W}; \boldsymbol{\tau})} (\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}_n) \right\|_2^2 \\
&\geq C_4 \left\| \boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}_n \right\|_2^2
\end{aligned}$$

where the second equality is due to the mean value theorem, and C_4 is the lower bound of $|q(\boldsymbol{W}; \boldsymbol{\tau})/2f^{\frac{1}{2}}(\boldsymbol{W}; \boldsymbol{\tau})|$. From the result of (A.3), it follows that $\|\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}_n\|_2^2 \to 0$ almost surely. Hence we have that $\rho(\hat{\boldsymbol{\pi}}_n, \boldsymbol{\pi}_0) \to 0$ almost surely as $n \to \infty$, which completes the proof. □

Proof of Theorem 2. To establish the convergence rate, note that by

DeVore et al. (1993) and DeVore (1998) and Lemma A1 of Lu, Zhang and Huang (2007), there exist h_{0nT} and h_{0nC} in $C^r[a,b]$ such that $\|h_{0nT} - h_{0T}^0\|_\infty = O_p\left(n^{-\zeta\nu}\right)$ and $\|h_{0nC} - h_{0C}^0\|_\infty = O_p\left(n^{-\zeta\nu}\right)$ for $0<\zeta \leq 1$, respectively.

For any $\epsilon > 0$, define the class

$$\mathcal{F}_\epsilon = \{\ell\left(\boldsymbol{\pi}_{0n}, \boldsymbol{W}\right) - \ell(\boldsymbol{\pi}, \boldsymbol{W}) : \boldsymbol{\pi} \in \Pi_n, d\left(\boldsymbol{\pi}, \boldsymbol{\pi}_{0n}\right) \leq \epsilon\}$$

$\rho(\hat{\boldsymbol{\pi}}_n, \boldsymbol{\xi}_0) \to 0$ with $\boldsymbol{\pi}_{0n} = (\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, h_{0nT}(x), h_{0nC}(x)) \in \Pi_n$. Following the calculation of Shen and Wong (1994) (p.597), denote $m_n = n^\nu$, we can establish that

$$\log N_{[]}\left(\varepsilon, \mathcal{F}_\epsilon, \|\cdot\|_2\right) \leq C m_n \log(\epsilon/\varepsilon)$$

Moreover, some algebraic calculations lead to $\|\ell\left(\boldsymbol{\pi}_{0n}, \boldsymbol{W}\right) - \ell(\boldsymbol{\pi}, \boldsymbol{W})\|_2^2 \leq C\epsilon^2$ for any $\ell\left(\boldsymbol{\pi}_{0n}, \boldsymbol{W}\right) - \ell(\boldsymbol{\pi}, \boldsymbol{W})$. Therefore, by Lemma 3.4.2 of van der Vaart and Wellner (1996) and Ma, Hu and Sun (2015) we obtain

$$E_P\|n^{1/2}(P_n\ell - P\ell)\|_{\mathcal{F}_\epsilon} \leq C J_\epsilon\left(\varepsilon, \mathcal{F}_\epsilon, \|\cdot\|_2\right)\left\{1 + \frac{J_\epsilon\left(\varepsilon, \mathcal{F}_\epsilon, \|\cdot\|_2\right)}{\epsilon^2 n^{1/2}}\right\}, \quad \text{(B.1)}$$

where

$$J_\epsilon\left(\varepsilon, \mathcal{F}_\epsilon, \|\cdot\|_2\right) = \int_0^\epsilon \left\{1 + \log N_{[]}\left(\varepsilon, \mathcal{F}_\epsilon, \|\cdot\|_2\right)\right\}^{1/2} d\varepsilon \leq C m_n^{1/2}\epsilon.$$

Hence we have

$$E_P\|n^{1/2}(P_n\ell - P\ell)\|_{\mathcal{F}_\epsilon} = O(m_n^{1/2}\epsilon + m_n/n^{1/2}).$$

Hence, the right-hand side of (B.1) yields $\phi_n(\epsilon) = C(m_n^{1/2}\epsilon + m_n/n^{1/2})$, which is the key function and covered in Theorem 3.2.5 of van der Vaart and Wellner (1996). It is easy to see that $\phi_n(\epsilon)/\epsilon$ is decreasing in ϵ, and

$$r_n^2\phi_n\left(1/r_n\right) = r_n m_n^{1/2} + r_n^2 m_n/n^{1/2} < 2n^{1/2},$$

where $r_n = m_n^{-1/2}n^{1/2} = n^{(1-\nu)/2}$ with $0 < \nu < 1/2$. Hence $n^{(1-\nu)/2}\rho(\hat{\boldsymbol{\pi}}_n, \boldsymbol{\pi}_{0n}) = O_p(1)$ by Theorem 3.4.1 of van der Vaart and Wellner (1996). Combining with Lemma A1 of Lu, Zhang and Huang (2007) and Ma, Hu and Sun (2015), we have

$$\begin{aligned}\rho(\boldsymbol{\pi}_{0n}, \boldsymbol{\pi}_0) =& \{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2\|_2^2 + \sup_{x\in[a,b]}\left|h_{0nT}(x) - h_{0T}^0(x)\right|^2 \\ &+ \sup_{x\in[a,b]}\left|h_{0nC}(x) - h_{0C}^0(x)\right|^2\}^{1/2} \\ &\leq c(\|h_{0nT}(x) - h_{0T}^0(x)\|_\infty + \|h_{0nC}(x) - h_{0C}^0(x)\|_\infty)\end{aligned}$$

Thus, $\rho(\boldsymbol{\pi}_{0n}, \boldsymbol{\pi}_0) = O_p(n^{-\zeta\nu})$, and this formulation yields that $\rho(\hat{\boldsymbol{\pi}}_n, \boldsymbol{\pi}_0) = O_p\left(n^{-(1-\nu)/2} + n^{-\zeta\nu}\right)$. The choice of $\nu = 1/(1+2\zeta)$ yields the rate of convergence $\rho(\hat{\boldsymbol{\pi}}_n, \boldsymbol{\pi}_0) = O_p\left(n^{-\zeta/(1+2\zeta)}\right)$. □

Lemma 1. *Assume that all conditions in Theorem 1 are satisfied, then with probability tending to 1, for any given* $\boldsymbol{\eta}_1$ *satisfying* $\|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_{10}\|_2 = O_p(n^{-1/2})$ *and any constant* C*, we have*

$$Q_n\left\{(\boldsymbol{\eta}_{11}^{\mathrm{T}}, 0)^{\mathrm{T}}\right\} = \max_{\|\boldsymbol{\eta}_{12}\| \le Cn^{-(1-\nu)/2}} Q_n\left\{(\boldsymbol{\eta}_{11}^{\mathrm{T}}, \boldsymbol{\eta}_{12}^{\mathrm{T}})^{\mathrm{T}}\right\}$$

where $Q_n(\boldsymbol{\eta}_1) = \frac{2}{n}\ell_n(\boldsymbol{\eta}_1) - \frac{1}{n}\sum_{j=1}^{2p} p_{mic}(|\eta_{1j}|)$ *and* $\ell_n(\boldsymbol{\eta}_1) = \ell(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2^{(k)})$ *with* $\boldsymbol{\eta}_2$ *fixed at its current estimate* $\boldsymbol{\eta}_2^{(k)}$*.*

Proof of Lemma 1. For any $\boldsymbol{\eta}_1$ satisfying $\|\boldsymbol{\eta}_{11} - \boldsymbol{\eta}_{10}\| = O_p(n^{-(1-\nu)/2})$ and any $\boldsymbol{\eta}_{12}$ satisfying $\|\boldsymbol{\eta}_{12}\| = O_p(n^{-1/2})$ in Assumption A5. It is sufficient to show that with probability tending to 1,

$$\begin{aligned} &\frac{\partial Q_n(\boldsymbol{\eta}_1)}{\partial \eta_{1j}} < 0 \quad \text{for} \quad 0 < \eta_{1j} < \varepsilon_n, \\ &\frac{\partial Q_n(\boldsymbol{\eta}_1)}{\partial \eta_{1j}} > 0 \quad \text{for} \quad -\varepsilon_n < \eta_{1j} < 0, \end{aligned} \tag{C.1}$$

as $n \to \infty$ and for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1, \ldots, 2p$. For $j = s+1, \ldots, 2p$, note that by the standard arguments,

$$\begin{aligned} &\frac{1}{n}\frac{\partial \ell_n(\boldsymbol{\eta}_{10})}{\partial \eta_{1j}} = O_p(n^{-1/2}) \\ &\frac{1}{n}\frac{\partial^2 \ell_n(\boldsymbol{\eta}_{10})}{\partial \eta_{1j}\partial \eta_{1l}} = E\left\{\frac{\partial^2 \ell_n(\boldsymbol{\eta}_{10})}{\partial \eta_{1j}\partial \eta_{1l}}\right\} + o_p(1) \end{aligned}$$

By Taylor's expansion and the assumption that $\|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_{10}\| = O_p(n^{-1/2})$,

we have

$$\begin{aligned}\frac{\partial Q_n(\boldsymbol{\eta}_1)}{\partial \eta_{1j}} &= \frac{2}{n}\frac{\partial \ell_n(\boldsymbol{\eta}_1)}{\partial \eta_{1j}} - \frac{1}{n}p'_{mic}(|\eta_{1j}|)\,\mathrm{sgn}(\eta_{1j}) \\ &= \frac{2}{n}\frac{\partial \ell_n(\boldsymbol{\eta}_{10})}{\partial \eta_{1j}} + \frac{2}{n}\sum_{l=1}^{2p}\frac{\partial^2 \ell_n(\boldsymbol{\eta}_{10})}{\partial \eta_{1j}\partial \eta_{1l}}(\eta_{1l} - \eta_{10,l}) \\ &\qquad + o_P(1) - \frac{1}{n}p'_{mic}(|\eta_{1j}|)\,\mathrm{sgn}(\eta_{1j}) \\ &= -\frac{1}{n}p'_{mic}(|\eta_{1j}|)\,\mathrm{sgn}(\eta_{1j}) + O_p(n^{-1/2}).\end{aligned} \tag{C.2}$$

For the first term in the formula above, note that $\eta_{1j} = O_p(n^{-1/2})$ yet $\eta_{1j} \neq 0$ for $\eta_{1j} \in \boldsymbol{\eta}_{12}$. Since $a_n = O_p(n)$ and $n_0 = O_p(n)$, similar arguments to Su et al. (2016) yield,

$$\sqrt{n}\cdot p'_{mic}(|\eta_{1j}|) = \frac{8\exp\left(2a\eta_{1j}^2\right)}{\{\exp\left(2a\eta_{10,j}^2\right)+1\}^2}a\ln(n_0)\sqrt{n}|\eta_{1j}| = O\left(\ln(n_0)n\right) \tag{C.3}$$

as $n \to \infty$. By (C.3), it implies that $p'_{mic}(|\eta_{1j}|) = O_p(\ln(n)\sqrt{n})$ as $n \to \infty$. Thus, by (C.2), it follows that

$$\begin{aligned}\frac{\partial Q_n(\boldsymbol{\eta}_1)}{\partial \eta_{1j}} &= -n^{-1}p'_{mic}(|\eta_{1j}|)\,\mathrm{sgn}\left(\eta_{1j}\right) + O_p(n^{-1/2}) \\ &= -n^{-/2}\ln(n)\,\mathrm{sgn}\left(\eta_{1j}\right) + O_p(n^{-1/2}),\end{aligned}$$

that is, the sign of the derivative $\partial Q_n(\boldsymbol{\eta}_1)/\partial \eta_{1j}$ is completely determined by the sign of η_{1j} when n is large, and they always have different signs. Hence, (C.1) follows. This completes the proof. □

Proof of Theorem 3. Similar to Xu et al. (2018), when implementing piecewise constant approximations to the baseline hazards, we often experience active constraints (i.e., some θ_u or γ_u are zero) when estimating $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ since the number and location of bins are not optimized; this can lead to singular information matrices. This fact has to be considered when developing asymptotic results. Let $\boldsymbol{\eta}_0$ represent the "true value" of parameter $\boldsymbol{\eta}$.

Let $\bar{\ell}(\boldsymbol{\eta}) = E_{\boldsymbol{\eta}_0}\left[n^{-1}\ell(\boldsymbol{\eta})\right]$. Based on the results of the strong law of large number that $n^{-1}\ell(\boldsymbol{\eta}) \to \bar{\ell}(\boldsymbol{\eta})$ almost surely and uniformly for $\boldsymbol{\eta} \in \Omega$, together with $\mu_n \to 0$ as $n \to \infty$ and $\boldsymbol{\eta}_0$ being the unique maximum of $\bar{\ell}(\boldsymbol{\eta})$ due to Assumption B2. This implies that the MPL estimates $\hat{\boldsymbol{\eta}} \to \boldsymbol{\eta}_0$ almost surely by applying, for example, Corollary 1 of Honoré and Powell (1994). Note that the consistency result requires that $K(\cdot,\cdot)$ is the correct copula function for T and C to obtain the correct log-likelihood function $\ell(\cdot)$.

It follows by Lemma 1 that part (i) holds. Now we prove part (ii). To prove the asymptotic normality result we need the following Taylor series expansion:

$$\frac{\partial \Phi_n(\widehat{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}} = \frac{\partial \Phi_n\left(\boldsymbol{\eta}_0\right)}{\partial \boldsymbol{\eta}} + \frac{\partial^2 \Phi_n(\widetilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top}\left(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\right)$$

where $\widetilde{\boldsymbol{\eta}}$ is a vector between $\widehat{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}_0$. This expansion requires that the function $\Phi_n(\boldsymbol{\eta})$ can be twice continuously differentiable in the neighborhood

of $\boldsymbol{\eta}_0$ as stated in Assumption B4. From the KKT conditions (see Section 2.3), we have that the constrained MPL estimate $\widehat{\boldsymbol{\eta}}$ satisfies

$$\boldsymbol{U}^\top \frac{\partial \Phi_n(\widehat{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}} = 0,$$

therefore

$$0 = \boldsymbol{U}^\top \frac{\partial \Phi_n\left(\boldsymbol{\eta}_0\right)}{\partial \boldsymbol{\eta}} + \boldsymbol{U}^\top \frac{\partial^2 \Phi_n(\widetilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top} \left(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\right).$$

Let $\widehat{\boldsymbol{\pi}}$ be $\widehat{\boldsymbol{\eta}}$ after deleting the active constraints and $\boldsymbol{\pi}_0$ defined similarly corresponding to $\boldsymbol{\eta}_0$, then

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 = \boldsymbol{U}\left(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0\right).$$

Thus we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\right) = -\boldsymbol{U}\left(\boldsymbol{U}^\top \frac{1}{n} \frac{\partial^2 \Phi_n(\tilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top} \boldsymbol{U}\right)^{-1} \boldsymbol{U}^\top \left(\frac{1}{\sqrt{n}} \frac{\partial \ell\left(\boldsymbol{\eta}_0\right)}{\partial \boldsymbol{\eta}} + o(1)\right).$$

Next, since $\boldsymbol{\eta}_0$ maximizes $E\left[n^{-1}\ell(\boldsymbol{\eta})\right]$, we have

$$E\left(\frac{\partial \ell\left(\boldsymbol{\eta}_0\right)}{\partial \boldsymbol{\eta}}\right) = 0$$

Thus, by the central limit theory, $n^{-1/2}\partial \ell\left(\boldsymbol{\eta}_0\right)/\partial \boldsymbol{\eta}$ converges in distribution to $N(0,\ I_0\left(\boldsymbol{\eta}_0\right))$. On the other hand,

$$-\frac{1}{n} \frac{\partial^2 \Phi_n(\widetilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \phi_i(\widetilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top},$$

which converges, according to the uniform strong law of large numbers and the consistency result, to $\boldsymbol{G}_0$ almost surely.

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \to N\left(\mathbf{0}_{2\times(m+p)}, \widetilde{\boldsymbol{G}}_0\left(\boldsymbol{\eta}_0\right)^{-1} \boldsymbol{I}_0\left(\boldsymbol{\eta}_0\right) \left[\widetilde{\boldsymbol{G}}_0\left(\boldsymbol{\eta}_0\right)^{-1}\right]^{\top}\right)$$

where $\widetilde{\boldsymbol{G}}_0\left(\boldsymbol{\eta}_0\right)^{-1} = \boldsymbol{U}\left(\boldsymbol{U}^{\top}\boldsymbol{G}_0(\boldsymbol{\eta})\boldsymbol{U}\right)^{-1}\boldsymbol{U}^{\top}$ and $\boldsymbol{I}_0\left(\boldsymbol{\eta}_0\right)$ is the expected information matrix. □

S2 Tuning parameter selection and algorithm implementation

Tuning parameter selection is a critical issue in most variable selection procedures. As a popular procedure, Generalized Cross-validation (GCV) has been extensively used to tune regularization parameters (Tibshirani, 1996; Fan and Li, 2001). However, Wang, Li and Tsai (2007) showed that the GCV approach tends to produce overfitted models if a finite-dimensional model truly exists. Hence, we follow Wang, Li and Tsai (2007) and Wang and Leng (2007) to consider an extended Bayesian Information Criterion (EBIC) as follows:

$$\text{EBIC}(a) = \ell(\widehat{\boldsymbol{\eta}}_1(a)) + \widehat{df}_{\sigma}\frac{\log(n)}{n},$$

where $\widehat{df}_a$ is the number of nonzero coefficients in $\hat{\boldsymbol{\eta}}_1(a)$, a simple estimate for the degrees of freedom (Zou, Hastie and Tibshirani, 2007). In this study, we propose estimating the degrees of freedom for the MIC estimator

by the number of selected coefficients: $\widehat{df_a} = \hat{s}$, where $\hat{s}$ is the cardinality of $\widehat{\mathscr{C}} = \{j : \widehat{\eta}_{1j} \neq 0, 1 \leq j \leq 2p\}$.

In Section 2.4, one can obtain the MIC estimator $\hat{\boldsymbol{\eta}}_1$ by solving (2.16). The resulting iterative algorithm (Algorithm 1) is presented as follows:

Algorithm 1

Notations:
L: the maximum iterative number.
$\boldsymbol{\eta}_1^{(k)}$: the k-th iteration estimate
$\hat{\boldsymbol{\eta}}_1$: the approximate solution to the quadratic minimization problem (2.12)
Step 1. Set the initial value $\boldsymbol{\eta}_1^{(0)} = 0$, and choose an initial step size t;
Step 2. For each $k \in \{1, 2, \ldots, L\}$, let $\tilde{\boldsymbol{\eta}}_1^{(k-1)} = \boldsymbol{\eta}_1^{(k-1)} - t^{-1}\ell'(\boldsymbol{\eta}_1^{(k-1)})$, then we have

$$\boldsymbol{\eta}_1^{(k)} = \{t\mathrm{I}_p + W_\lambda(\boldsymbol{\eta}_1^{(k-1)})\}^{-1}\{t\tilde{\boldsymbol{\eta}}_1^{(k-1)}\},$$

where $W_\lambda(\boldsymbol{\eta}_1^{(k-1)}) = \mathrm{diag}\left\{p'_{mic}(|\eta_{11}^{(k-1)}|)/|\eta_{11}^{(k-1)}|, \cdots, p'_{mic}(|\eta_{1,2p}^{(k-1)}|)/|\eta_{1,2p}^{(k-1)}|\right\}$.
Step 3. Stop the algorithm until

$$\|\boldsymbol{\eta}_1^{(k)} - \boldsymbol{\eta}_1^{(k-1)}\|_2^2 \leq \varepsilon\|\boldsymbol{\eta}_1^{(k-1)}\|_2^2.$$

Thus, $\hat{\boldsymbol{\eta}}_1 = \boldsymbol{\eta}_1^{(k)}$ is an optimal solution to (2.16); Otherwise, set $k = k + 1$ and return to Step 2.

S3 Additional numerical studies

S3.1 Robustness of MPL regarding p

We consider $p = 60$ and 100 to investigate the performance of the proposed MPL when p is relatively large. We simulate 500 datasets using the same setting as in the previous simulation under $n = 200$ and 400 and compute the five measures, Pcorr, MSE, F_+, F_-, and Size. Tables S2 and S3

summarize the MSE and variable selection results in Cases A and B, respectively. Tables S4 presents the simulation results of the bias (BIAS), the sample standard errors (SE), the average of the asymptotic standard errors (ASE), and the coverage probabilities (CP) of the 95% confidence intervals for nonzero coefficients obtained by the MPL proposed and conventional PL methods in Case A under dependent censoring. As shown in Tables S2 and S3, our method still performs satisfactorily when p is relatively large and outperforms the full model and PL method in terms of Pcorr and MSE and performs similarly to the Oracle model. Between the two procedures with penalties, the MPL performs consistently better than PL in terms of all the performance measures, and the performance of both methods improves in terms of all the measures as the sample size n increases from 200 to 400 but declines when the model size p rises from 60 to 100. All results are consistent with the previous simulation results in Tables 1–3.

S3.2 Comparison with model selection on $\boldsymbol{\beta}$ only

From Equation (2.5), model selection was performed on $\boldsymbol{\eta}_1$, including both $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$. Since $\boldsymbol{\phi}$ is only related to censored data, we conduct a simulation to examine the necessity of selecting $\boldsymbol{\phi}$. We generated 500 datasets under Case A with CR = 30% and $n = 200$ or 400. We compute the bias (BIAS),

the sample standard errors (SE), the average of the asymptotic standard errors (ASE), and the coverage probabilities (CP) of the 95% confidence intervals for nonzero coefficients obtained by the proposed and PL methods under dependent censoring. Table S5 shows that the proposed MPL keeps a better agreement between SE and ASE values than the MPL* (only select $\boldsymbol{\beta}$), and the coverage probabilities of the 95% confidence intervals yielded by our method are closer to the nominal level than the MPL* method. Thus, selecting $\boldsymbol{\phi}$ helps improve the model selection results on $\boldsymbol{\beta}$.

S3.3 Comparison with other algorithm based on M-spline

This simulation examines model selection results obtained by the MPL methods with piecewise constant and spline basis functions. We consider the settings of $p = 10, 20$ and $\tau = 0.3, 0.5$. Table S6 presents the comparison results, where MPL-MS denotes MPL with M-spline basis functions for the baseline hazard. As shown in Table S6, both methods perform comparably. In particular, our method outperforms MPL-MS in MSE when $\tau = 0.3$ and underperforms MPL-MS when $\tau = 0.5$.

Table S1: Simulation results of MPL, PL, full model, and Oracle model in Case A under independent censoring, Frank copula, and $\tilde{\tau} = 0.5$

			Censoring rate 25%					Censoring rate 45%				
n	(ρ, p)	Method	Pcorr	MSE	F_+	F_-	Size	Pcorr	MSE	F_+	F_-	Size
$n = 200$	$(0.25, 10)$	Full	0.00	0.0837	6.994	0.000	9.994	0.00	0.1160	6.992	0.000	9.992
		PL	95.80	0.0299	0.050	0.046	3.004	90.60	0.0447	0.018	0.000	3.018
		MPL	95.80	0.0416	0.020	0.024	2.996	88.40	0.0567	0.040	0.082	2.958
		Oracle	100.00	0.0233	0.000	0.000	3.000	100.00	0.0303	0.000	0.000	3.000
	$(0.25, 20)$	Full	0.00	0.2027	16.984	0.000	19.984	0.00	0.2724	16.994	0.000	19.994
		PL	94.40	0.0359	0.036	0.024	3.012	84.40	0.0608	0.120	0.064	3.056
		MPL	93.80	0.0479	0.034	0.034	3.000	84.40	0.0669	0.070	0.102	2.968
		Oracle	100.00	0.0257	0.000	0.000	3.000	100.00	0.0322	0.000	0.000	3.000
	$(0.5, 10)$	Full	0.00	0.0820	6.996	0.000	9.996	0.00	0.1078	6.998	0.000	9.998
		PL	91.40	0.0332	0.060	0.032	3.028	83.40	0.0527	0.132	0.070	3.062
		MPL	91.60	0.0431	0.054	0.040	3.014	82.40	0.0649	0.110	0.110	3.000
		Oracle	100.00	0.0227	0.000	0.000	3.000	100.00	0.0306	0.000	0.000	3.000
	$(0.5, 20)$	Full	0.00	0.1982	16.994	0.000	19.994	0.00	0.2626	16.982	0.000	19.982
		PL	88.00	0.0415	0.124	0.044	3.080	75.40	0.0686	0.294	0.072	3.222
		MPL	87.40	0.0537	0.132	0.046	3.086	76.40	0.0762	0.244	0.098	3.146
		Oracle	100.00	0.0246	0.000	0.000	3.000	100.00	0.0316	0.000	0.000	3.000
$n = 400$	$(0.25, 10)$	Full	0.00	0.0381	6.990	0.000	9.990	0.00	0.0495	6.996	0.000	9.996
		PL	99.80	0.0118	0.002	0.000	3.002	99.60	0.0146	0.002	0.002	3.000
		MPL	99.80	0.0211	0.002	0.000	3.002	99.20	0.0193	0.002	0.006	2.996
		Oracle	100.00	0.0116	0.000	0.000	3.000	100.00	0.0138	0.000	0.000	3.000
	$(0.25, 20)$	Full	0.00	0.0817	16.974	0.000	19.974	0.00	0.1092	16.982	0.000	19.982
		PL	100.00	0.0127	0.000	0.000	3.000	99.60	0.0172	0.000	0.004	2.996
		MPL	100.00	0.0232	0.000	0.000	3.000	99.20	0.0213	0.000	0.010	2.990
		Oracle	100.00	0.0122	0.000	0.000	3.000	100.00	0.0158	0.000	0.000	3.000
	$(0.5, 10)$	Full	0.00	0.0374	6.990	0.000	9.990	0.00	0.0488	6.992	0.000	9.992
		PL	99.80	0.0116	0.000	0.002	2.998	98.40	0.0167	0.004	0.014	2.990
		MPL	99.80	0.0206	0.000	0.002	2.998	97.60	0.0224	0.004	0.022	2.982
		Oracle	100.00	0.0112	0.000	0.000	3.000	100.00	0.0137	0.000	0.000	3.000
	$(0.5, 20)$	Full	0.00	0.0795	16.974	0.000	19.974	0.00	0.1062	16.992	0.000	19.992
		PL	100.00	0.0122	0.000	0.000	3.000	98.40	0.0186	0.006	0.014	2.992
		MPL	100.00	0.0221	0.000	0.000	3.000	97.80	0.0225	0.010	0.016	2.994
		Oracle	100.00	0.0119	0.000	0.000	3.000	100.00	0.0153	0.000	0.000	3.000

Table S2: Simulation results of MPL, PL, full model, and Oracle model in Case A (CR=30%)

			Frank copula and $\tilde{\tau} = 0.2$, $\lambda_t = 2.0$					Frank copula and $\tilde{\tau} = 0.5$, $\lambda_t = 2.3$				
n	(ρ, p)	Method	Pcorr	MSE	F_+	F_-	Size	Pcorr	MSE	F_+	F_-	Size
$n = 200$	$(0.25, 60)$	Full	0.00	1.000	56.926	0.000	59.926	0.00	1.108	56.940	0.000	59.940
		PL	86.20	0.048	0.136	0.020	3.116	82.20	0.079	0.200	0.004	3.196
		MPL	100.00	0.026	0.000	0.000	3.000	99.80	0.068	0.000	0.002	2.998
		Oracle	100.00	0.024	0.000	0.000	3.000	100.00	0.038	0.000	0.000	3.000
	$(0.25, 100)$	Full	0.00	2.786	96.892	0.000	99.892	0.00	2.974	96.898	0.000	99.898
		PL	67.40	0.072	0.400	0.028	3.372	68.00	0.097	0.392	0.008	3.384
		MPL	100.00	0.030	0.000	0.000	3.000	100.00	0.076	0.000	0.000	3.000
		Oracle	100.00	0.024	0.000	0.000	3.000	100.00	0.035	0.000	0.000	3.000
	$(0.5, 60)$	Full	0.00	0.867	56.934	0.000	59.934	0.00	0.949	56.936	0.000	59.936
		PL	83.40	0.051	0.132	0.058	3.074	84.40	0.070	0.168	0.010	3.158
		MPL	100.00	0.024	0.000	0.000	3.000	99.80	0.058	0.000	0.002	2.998
		Oracle	100.00	0.023	0.000	0.000	3.000	100.00	0.035	0.000	0.000	3.000
	$(0.5, 100)$	Full	0.00	2.284	96.902	0.000	99.902	0.00	2.409	96.904	0.000	99.904
		PL	70.60	0.069	0.300	0.076	3.224	73.60	0.084	0.320	0.016	3.304
		MPL	100.00	0.026	0.000	0.000	3.000	100.00	0.063	0.000	0.000	3.000
		Oracle	100.00	0.024	0.000	0.000	3.000	100.00	0.034	0.000	0.000	3.000
$n = 400$	$(0.25, 60)$	Full	0.00	0.315	56.886	0.000	59.886	0.00	0.376	56.874	0.000	59.874
		PL	100.00	0.013	0.000	0.000	3.000	93.80	0.061	0.064	0.000	3.064
		MPL	100.00	0.012	0.000	0.000	3.000	100.00	0.040	0.000	0.000	3.000
		Oracle	100.00	0.011	0.000	0.000	3.000	100.00	0.021	0.000	0.000	3.000
	$(0.25, 100)$	Full	0.00	0.725	96.818	0.000	99.818	0.00	0.823	96.830	0.000	99.830
		PL	99.40	0.015	0.004	0.002	3.002	95.40	0.064	0.046	0.000	3.046
		MPL	100.00	0.014	0.000	0.000	3.000	100.00	0.049	0.000	0.000	3.000
		Oracle	100.00	0.011	0.000	0.000	3.000	100.00	0.023	0.000	0.000	3.000
	$(0.5, 60)$	Full	0.00	0.283	56.900	0.000	59.900	0.00	0.331	56.886	0.000	59.886
		PL	98.00	0.018	0.022	0.000	3.022	96.20	0.043	0.042	0.000	3.042
		MPL	100.00	0.012	0.000	0.000	3.000	100.00	0.043	0.000	0.000	3.000
		Oracle	100.00	0.011	0.000	0.000	3.000	100.00	0.020	0.000	0.000	3.000
	$(0.5, 100)$	Full	0.00	0.635	96.848	0.000	99.848	0.00	0.708	96.840	0.000	99.840
		PL	97.40	0.022	0.034	0.000	3.034	89.40	0.063	0.120	0.000	3.120
		MPL	100.00	0.014	0.000	0.000	3.000	100.00	0.047	0.000	0.000	3.000
		Oracle	100.00	0.012	0.000	0.000	3.000	100.00	0.021	0.000	0.000	3.000

Table S3: Simulation results of MPL, PL, full model, and Oracle model in Case B (CR=30%)

			Frank copula and $\tilde{\tau}=0.2$, $\lambda_t=1.5$					Frank copula and $\tilde{\tau}=0.5$, $\lambda_t=1.6$				
n	(ρ, p)	Method	Pcorr	MSE	F_+	F_-	Size	Pcorr	MSE	F_+	F_-	Size
$n=200$	$(0.25, 60)$	Full	0.00	1.482	56.950	0.000	59.950	0.00	1.025	56.932	0.000	59.932
		PL	80.20	0.058	0.252	0.000	3.252	80.40	0.080	0.232	0.000	3.232
		MPL	100.00	0.035	0.000	0.000	3.000	100.00	0.036	0.000	0.000	3.000
		Oracle	100.00	0.035	0.000	0.000	3.000	100.00	0.030	0.000	0.000	3.000
	$(0.25, 100)$	Full	0.00	5.383	96.912	0.000	99.912	0.00	2.906	96.918	0.000	99.918
		PL	56.80	0.092	0.650	0.000	3.650	59.00	0.122	0.620	0.000	3.620
		MPL	100.00	0.076	0.000	0.000	3.000	100.00	0.055	0.000	0.000	3.000
		Oracle	100.00	0.033	0.000	0.000	3.000	100.00	0.028	0.000	0.000	3.000
	$(0.5, 60)$	Full	0.00	1.349	56.938	0.000	59.938	0.00	1.478	56.946	0.000	59.946
		PL	81.00	0.055	0.256	0.004	3.256	74.60	0.814	0.320	0.000	3.320
		MPL	100.00	0.036	0.000	0.000	3.000	100.00	0.068	0.000	0.000	3.000
		Oracle	100.00	0.033	0.000	0.000	3.000	100.00	0.051	0.000	0.000	3.000
	$(0.5, 100)$	Full	0.00	4.482	96.934	0.000	99.934	0.00	4.758	96.950	0.000	99.950
		PL	61.00	0.084	0.646	0.004	3.646	63.40	0.108	0.624	0.000	3.624
		MPL	100.00	0.058	0.000	0.000	3.000	100.00	0.106	0.000	0.000	3.000
		Oracle	100.00	0.033	0.000	0.000	3.000	100.00	0.048	0.000	0.000	3.000
$n=400$	$(0.25, 60)$	Full	0.00	0.393	56.910	0.000	59.910	0.00	0.507	56.870	0.000	59.870
		PL	98.20	0.040	0.018	0.000	3.018	89.40	0.116	0.120	0.000	3.120
		MPL	100.00	0.020	0.000	0.000	3.000	100.00	0.097	0.000	0.000	3.000
		Oracle	100.00	0.015	0.000	0.000	3.000	100.00	0.028	0.000	0.000	3.000
	$(0.25, 100)$	Full	0.00	0.984	96.838	0.000	99.838	0.00	1.153	96.860	0.000	99.860
		PL	96.40	0.054	0.036	0.000	3.036	88.20	0.140	0.128	0.000	3.128
		MPL	100.00	0.028	0.000	0.000	3.000	100.00	0.075	0.000	0.000	3.000
		Oracle	100.00	0.016	0.000	0.000	3.000	100.00	0.030	0.000	0.000	3.000
	$(0.5, 60)$	Full	0.00	0.372	56.902	0.000	59.902	0.00	0.452	56.916	0.000	59.916
		PL	96.00	0.036	0.042	0.000	3.042	84.40	0.101	0.186	0.000	3.186
		MPL	100.00	0.018	0.000	0.000	3.000	100.00	0.082	0.000	0.000	3.000
		Oracle	100.00	0.014	0.000	0.000	3.000	100.00	0.027	0.000	0.000	3.000
	$(0.5, 100)$	Full	0.00	0.907	96.850	0.000	99.850	0.00	1.032	96.892	0.000	99.892
		PL	94.60	0.051	0.062	0.004	3.062	80.40	0.136	0.240	0.000	3.240
		MPL	100.00	0.027	0.000	0.000	3.000	100.00	0.059	0.000	0.000	3.000
		Oracle	100.00	0.015	0.000	0.000	3.000	100.00	0.030	0.000	0.000	3.000

Table S4: Simulation results of MPL and PL in Case A under dependent censoring (CR=30%, $\lambda_t = 2.0$), Frank copula, and $\tilde{\tau} = 0.2$. Abbreviations: SE, the sample standard error; ASE, the averages of asymptotic standard error; BIAS, bias; CP, the coverage probability of the 95% confidence intervals for the nonzero coefficients.

			MPL				PL			
n	(ρ, p)	Para	BIAS	SE	ASE	CP	BIAS	SE	ASE	CP
$n = 200$	$(0.25, 60)$	β_1	0.014	0.057	0.061	0.956	0.014	0.112	0.037	0.492
		β_3	-0.034	0.058	0.061	0.918	-0.030	0.117	0.035	0.456
		β_6	0.040	0.055	0.061	0.918	0.038	0.103	0.034	0.450
		ϕ_2	-0.021	0.076	0.086	0.986	- -	- -	- -	- -
		ϕ_4	0.019	0.083	0.085	0.984	- -	- -	- -	- -
	$(0.25, 100)$	β_1	0.018	0.051	0.061	0.968	0.019	0.117	0.036	0.500
		β_3	-0.025	0.054	0.061	0.962	-0.024	0.120	0.033	0.436
		β_6	0.036	0.057	0.060	0.914	0.032	0.116	0.033	0.444
		ϕ_2	0.028	0.090	0.086	0.958	- -	- -	- -	- -
		ϕ_4	0.022	0.079	0.086	0.972	- -	- -	- -	- -
	$(0.5, 60)$	β_1	0.008	0.052	0.062	0.974	0.013	0.119	0.037	0.484
		β_3	-0.032	0.053	0.063	0.952	-0.029	0.126	0.033	0.424
		β_6	0.046	0.051	0.061	0.920	-0.040	0.111	0.029	0.380
		ϕ_2	0.023	0.069	0.088	0.984	- -	- -	- -	- -
		ϕ_4	-0.020	0.074	0.088	0.988	- -	- -	- -	- -
	$(0.5, 100)$	β_1	-0.008	0.054	0.062	0.960	-0.014	0.153	0.039	0.480
		β_3	-0.004	0.054	0.063	0.974	-0.012	0.170	0.037	0.406
		β_6	0.032	0.060	0.061	0.902	0.016	0.151	0.031	0.370
		ϕ_2	0.038	0.081	0.089	0.976	- -	- -	- -	- -
		ϕ_4	-0.029	0.076	0.088	0.982	- -	- -	- -	- -
$n = 400$	$(0.25, 60)$	β_1	0.012	0.031	0.042	0.976	0.014	0.061	0.027	0.598
		β_3	-0.025	0.033	0.042	0.952	-0.023	0.065	0.026	0.526
		β_6	0.034	0.032	0.042	0.926	0.031	0.063	0.026	0.512
		ϕ_2	-0.018	0.046	0.059	0.952	- -	- -	- -	- -
		ϕ_4	0.018	0.044	0.059	0.962	- -	- -	- -	- -
	$(0.25, 100)$	β_1	0.019	0.030	0.042	0.984	0.020	0.066	0.027	0.574
		β_3	-0.029	0.032	0.042	0.948	-0.028	0.066	0.025	0.498
		β_6	0.035	0.032	0.042	0.918	0.033	0.066	0.025	0.536
		ϕ_2	-0.019	0.055	0.059	0.958	- -	- -	- -	- -
		ϕ_4	0.020	0.049	0.059	0.952	- -	- -	- -	- -
	$(0.5, 60)$	β_1	0.003	0.027	0.043	0.994	0.005	0.063	0.029	0.598
		β_3	-0.021	0.030	0.044	0.972	-0.017	0.070	0.026	0.524
		β_6	0.041	0.028	0.043	0.930	0.036	0.062	0.024	0.480
		ϕ_2	-0.018	0.039	0.061	0.968	- -	- -	- -	- -
		ϕ_4	0.016	0.038	0.061	0.972	- -	- -	- -	- -
	$(0.5, 100)$	β_1	0.007	0.028	0.043	0.988	0.006	0.079	0.030	0.602
		β_3	-0.020	0.030	0.044	0.974	-0.015	0.092	0.025	0.490
		β_6	0.040	0.030	0.042	0.912	0.032	0.089	0.024	0.474
		ϕ_2	-0.014	0.051	0.061	0.966	- -	- -	- -	- -
		ϕ_4	0.013	0.045	0.061	0.970	- -	- -	- -	- -

Table S5: Simulation results of MPL and MPL* in Case A under dependent censoring (CR=30%, $\lambda_t = 2.3$), Frank copula, and $\tilde{\tau} = 0.5$. Abbreviations: MPL*, the special case for MPL where the model selection was only performed on $\boldsymbol{\beta}$; SE, the sample standard error; ASE, the averages of asymptotic standard error; BIAS, bias; CP, the coverage probability of the 95% confidence intervals for the nonzero coefficients.

			MPL (select $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$)				MPL* (only select $\boldsymbol{\beta}$)			
n	(ρ, p)	Para	BIAS	SE	ASE	CP	BIAS	SE	ASE	CP
$n = 200$	$(0.25, 20)$	β_1	-0.002	0.044	0.050	0.974	0.017	0.061	0.050	0.860
		β_3	-0.012	0.050	0.050	0.922	-0.032	0.065	0.050	0.828
		β_6	0.018	0.045	0.050	0.942	0.039	0.061	0.050	0.814
	$(0.25, 60)$	β_1	0.002	0.044	0.050	0.966	0.023	0.075	0.050	0.792
		β_3	-0.024	0.043	0.050	0.950	-0.053	0.076	0.051	0.716
		β_6	0.035	0.043	0.050	0.920	0.063	0.072	0.050	0.664
	$(0.25, 100)$	β_1	0.013	0.040	0.051	0.972	0.035	0.073	0.051	0.784
		β_3	-0.027	0.043	0.051	0.954	-0.054	0.079	0.051	0.686
		β_6	0.041	0.045	0.051	0.904	0.062	0.082	0.051	0.642
	$(0.5, 20)$	β_1	-0.026	0.044	0.050	0.956	-0.003	0.066	0.050	0.856
		β_3	0.017	0.051	0.051	0.940	-0.020	0.071	0.051	0.818
		β_6	0.015	0.048	0.050	0.940	0.053	0.065	0.050	0.720
	$(0.5, 60)$	β_1	-0.027	0.044	0.051	0.958	0.001	0.078	0.051	0.796
		β_3	-0.005	0.047	0.051	0.960	-0.041	0.078	0.051	0.740
		β_6	0.029	0.047	0.050	0.916	-0.075	0.075	0.050	0.610
	$(0.5, 100)$	β_1	-0.023	0.036	0.051	0.984	0.006	0.071	0.052	0.858
		β_3	-0.012	0.039	0.052	0.980	-0.035	0.074	0.052	0.788
		β_6	0.027	0.044	0.050	0.934	0.070	0.081	0.051	0.592
$n = 400$	$(0.25, 20)$	β_1	-0.006	0.033	0.035	0.954	0.006	0.039	0.035	0.908
		β_3	-0.007	0.031	0.035	0.956	-0.021	0.037	0.035	0.892
		β_6	0.012	0.031	0.035	0.954	0.026	0.038	0.035	0.856
	$(0.25, 60)$	β_1	0.011	0.024	0.034	0.980	0.024	0.043	0.035	0.844
		β_3	-0.021	0.026	0.035	0.968	-0.042	0.044	0.035	0.704
		β_6	0.031	0.024	0.035	0.934	0.041	0.042	0.035	0.638
	$(0.25, 100)$	β_1	0.022	0.019	0.034	0.978	0.034	0.043	0.035	0.754
		β_3	-0.028	0.020	0.035	0.974	-0.049	0.045	0.035	0.632
		β_6	0.038	0.021	0.035	0.920	0.059	0.045	0.035	0.576
	$(0.5, 20)$	β_1	-0.032	0.031	0.035	0.912	0.017	0.039	0.035	0.898
		β_3	0.018	0.031	0.036	0.962	-0.004	0.040	0.036	0.920
		β_6	0.010	0.034	0.035	0.942	0.034	0.039	0.035	0.800
	$(0.5, 60)$	β_1	-0.037	0.025	0.035	0.908	-0.014	0.035	0.036	0.862
		β_3	0.029	0.028	0.036	0.958	-0.013	0.036	0.036	0.854
		β_6	0.006	0.028	0.035	0.978	0.052	0.035	0.035	0.668
	$(0.5, 100)$	β_1	-0.031	0.025	0.035	0.944	-0.005	0.047	0.035	0.868
		β_3	0.025	0.026	0.036	0.964	-0.021	0.049	0.036	0.802
		β_6	0.011	0.027	0.035	0.964	0.061	0.048	0.035	0.582

Table S6: Results of MPL and MPL-MS in Case A under $n = 200$ and CR=30%. MPL-MS denotes MPL with M-spline basis function for the baseline hazard.

		Frank copula and $\tilde{\tau} = 0.3$, $\lambda_t = 2.1$					Frank copula and $\tilde{\tau} = 0.5$, $\lambda_t = 2.3$				
(ρ, p)	Method	Pcorr	MSE	F_+	F_-	Size	Pcorr	MSE	F_+	F_-	Size
$(0.25, 10)$	MPL	100.00	0.0084	0.000	0.000	3.000	100.00	0.0243	0.000	0.000	3.000
	MPL-MS	100.00	0.0365	0.000	0.000	3.000	100.00	0.0261	0.000	0.000	3.000
$(0.25, 20)$	MPL	100.00	0.0108	0.000	0.000	3.000	100.00	0.0333	0.000	0.000	3.000
	MPL-MS	100.00	0.0245	0.000	0.000	3.000	100.00	0.0086	0.000	0.000	3.000
$(0.5, 10)$	MPL	100.00	0.0079	0.000	0.000	3.000	100.00	0.0232	0.000	0.000	3.000
	MPL-MS	100.00	0.0227	0.000	0.000	3.000	100.00	0.0124	0.000	0.000	3.000
$(0.5, 20)$	MPL	100.00	0.0093	0.000	0.000	3.000	100.00	0.0277	0.000	0.000	3.000
	MPL-MS	100.00	0.0104	0.000	0.000	3.000	100.00	0.0092	0.000	0.000	3.000

Bibliography

DeVore, R. A. (1998). Nonlinear approximation. *Acta Numer.* **7**, 51–150.

DeVore, R. A., Kyriazis, G., Leviatan, D. and Tikhomirov, V. M. (1993). Wavelet compression and nonlinear n-widths. *Adv. Comput. Math.* **1**, 197–214.

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

Honoré, B. E. and Powell, J. L. (1994). Pairwise difference estimators of censored and truncated regression models. *J. Econometrics* **64**, 241–278.

Lu, M., Zhang, Y. and Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* **94**, 705–718.

Ma, L., Hu, T. and Sun, J. (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* **102**, 731–738.

Ma, L., Hu, T. and Sun, J. (2016). Cox regression analysis of dependent interval-censored failure time data. *Comput. Statist. Data Anal.* **103**, 79–90.

Pollard, D. (1984). *Convergence of Stochastic Processes.* 1st Edition. Springer New York.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580–615.

Su, X. G., Wijayasinghe, C. S., Fan, J. J. and Zhang, Y. (2016). Sparse estimation of cox proportional hazards models via approximated information criteria. *Biometrics* **72**, 751–

759.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer Berlin.

Wang, H. and Leng, C. (2007). Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc.* **102**, 1039–1048.

Wang, H., Li, R. and Tsai, C. (2007). On the consistency of scad tuning parameter selector. *Biometrika.* **94**, 553–568.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23**, 339–362.

Xu, J., Ma, J., Connors, M. H. and Brodaty, H. (2018). Proportional hazard model estimation under dependent censoring using copulas and penalized likelihood. *Stat. Med.* **37**, 2238–2251.

Xue, H., Lam, K. F. and Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *J. Amer. Statist. Assoc.* **99**, 346–356.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Ann. Statist.* **35**, 2173–2192.