Statistica Sinica Preprint No: SS-2023-0227

Title	Simultaneous Variable Selection and Estimation of Survival Model with Informative Censoring
Manuscript ID	SS-2023-0227
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0227
Complete List of Authors	Zili Liu, Hong Wang, Chunjie Wang and Xinyuan Song
Corresponding Authors	Xinyuan Song
E-mails	xysong@sta.cuhk.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

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

Abstract: This study proposes a maximum penalized likelihood procedure for simultaneous estimation and variable selection in the context of Cox proportional hazards models with informative right-censored data. A copula function is adopted to model the dependence between censoring and event times. Moreover, two penalty functions are introduced to accommodate the sparsity of regression coefficients and smooth the baseline hazard estimate. Since the baseline hazard function is nonnegative, we propose a specific algorithm comprising a modified Newton algorithm for updating regression coefficients and a multiplicative iterative algorithm for updating baseline hazard at each iteration. Furthermore, we establish the asymptotic properties of the proposed estimators. Simulation studies show that the proposed method performs satisfactorily. Finally, we apply the proposed method to investigate the potential risk factors of AIDS for HIV-1-infected patients from the AIDS Clinical Trials Group Protocol 175 study.

Key words and phrases: copulas, informative censoring, maximum penalized likelihood, variable selection.

1. Introduction

Survival data with informative censoring are commonly encountered in biomedical studies Examples include the Prospective Research In MEmory (PRIME) study (Brodaty et al., 2011) and the Acquired Immunodeficiency Syndrome (AIDS) Clinical Trials Group Protocol 175 (ACTG 175) study (Hammer et al., 1996). The PRIME concerned patients with either dementia or mild cognitive impairment, with time to institutionalization as the endpoint. Brodaty et al. (2014) showed that patients who withdrew from the study were older, had lower cognitive and functional abilities, and more severe neuropsychiatric symptoms, and were thus more likely to be institutionalized than those who remained in the study, suggesting a clear dependence between the withdrawal and endpoint times. The ACTG 175 study evaluated nucleoside monotherapy versus combination therapy in HIV-1 infected patients, with time to 50% decline in CD4 from baseline as the primary endpoint. Scharfstein and Robins (2002) showed that those reporting injection-drug use and with lower CD4 cell counts, lower Karnofsky scores, and symptoms of HIV infection at enrolment were significantly more likely to discontinue treatment before the study ended, suggesting informative censoring. They also revealed that the Kaplan-Meier estimators would overestimate the true treatment-specific survival curves when disregarding informative censoring as noninformative. An essential feature of informative censoring data is that the event time T and censoring time C are correlated, and ignoring such correction may yield unreliable results.

Meanwhile, variable selection for improving model efficiency and interpretability has received wide attention in survival analysis. Over the past decades, sparse estimation via a regularized or penalized log-likelihood (or estimating function) has attracted increasing interest. Commonly used penalized approaches include the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), LASSO-type estimator and its sparsity for penalized linear regression (Knight and Fu, 2000), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive LASSO (Zou, 2006), grouped LASSO (Yuan and Lin, 2006), the minimax concave penalty (MCP) (Zhang, 2010), and the seamless-L_0 (SELO) (Dicker, Huang and Lin, 2013), among others. In addition, various computational methods, such as the shooting method for implementing LASSO in the context of linear regression models (Fu, 1998), active shooting for partial correlation estimation based on joint sparse regression (Peng et al., 2009), least angle regression (LARS) for linear model selection (Efron et al., 2004), and the generalized cross-validation (GCV) procedure for tuning parameter estimation (Fu, 2005), have also been developed. Moreover, variable selection has attracted significant attention in survival analysis (see, e.g., Tibshirani, 1997; Fan and Li, 2002; Zhang and Lu, 2007; Zhao et al., 2020; Zhang et al., 2024, and references therein). Despite the fruitful literature mentioned above, existing works mainly focused on independent censoring. When informative censoring occurs, variable selection becomes highly challenging, and available state-of-the-art procedures are still lacking.

Many regression-based estimation approaches have been available for Cox proportional hazard (PH) models in the presence of informative right-censored data. Existing literature for PH models can be grouped into copula-based (Huang and Zhang, 2008; Chen, 2010; Chen, Hu and Sun, 2017; Xu et al., 2018; Jo et al., 2023) and frailty-based methods (Huang and Wolfe, 2002; Ha et al., 2014; Reeder, Lu and Haneuse, 2023). Frailty methods assume conditional independence between the failure and censoring times given a frailty term and focus only on estimating the conditional hazard function. For example, Xu et al. (2018) proposed maximum penalized likelihood (MPL) estimation for PH models under informative right-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

censored data. They estimated parameters by maximizing the nonconcave likelihood function, which involves estimating nonparametric functions and complicated gradients of the objective function. Hence, their method inevitably computed the inverse of the Hessian matrix or the second-order differential matrix with respect to regression parameters, getting into the dilemma of dealing with the singularity of these inverse matrices when the number of covariates becomes large. Unlike frailty methods, copula methods assess the marginal hazard functions, in which a copula function provides an approximated joint survival function of the failure and censoring times and facilitates estimation of the marginal hazard functions. However, due to the complexity of the data structure, a correlation between the censoring and survival times formulated by copulas or frailty terms, and the nonparametric baseline hazard function, variable selection by commonly used regularization approaches in the presence of informative censoring faces both sophisticated mathematical derivation and challenging implementation, making the computation and inference a formidable task.

To tackle these issues, we adopt the "minimum approximated information criterion (MIC)" (Su et al., 2016) to conduct variable selection for PH models with dependent right-censored data. Specifically, we propose a novel variable selection procedure based on MPL with a MIC penalty.

Motivated by Xu et al. (2018), an assumed copula function is adopted to model the dependence between censoring and event times, and two penalty functions are introduced to encourage the sparsity of the regression coefficients and smooth the baseline hazard estimate, respectively. We develop a modified Newton multiplicative-iterative algorithm to estimate the baseline hazard function and regression coefficients. Compared to Xu et al. (2018), the proposed algorithm is computationally more efficient because it can avoid computing the inverse of the Hessian matrix and thus considerably save computational costs in the case of large-scale covariates. Although coordinate-descent approaches, such as the shooting algorithm (Fu, 1998) and LARS (Efron et al., 2004; Peng et al., 2009), can also avoid computing the inverse matrix and have been demonstrated efficient in analyzing independently censored data, directly applying them in the presence of informative censoring is challenging because of the highly intractable likelihood induced by the complex data structure, the presence of correlation between the censoring and survival times described by copula functions, and the nonparametric approximation of the basis functions. Unlike available methods in the literature, the proposed algorithm approximates the nonconcave likelihood function by a new convex objective function, which enables us to have a closed-form solution for the optimization problem involving highly intractable likelihood function, thereby performing stably and efficiently. We also establish the asymptotic properties of these estimates and provide their convergence rate under mild conditions.

The remainder of the article is organized as follows. Section 2 presents the penalized log-likelihood function based on an assumed copula and the MPL estimation procedure for regression parameters. Section 2 discusses how to compute the MPL estimate of the nonnegatively constrained baseline hazard and the regression coefficients. We propose a BIC-type tuning parameter selection method for the MIC procedure in the Supplementary Material. The asymptotic properties of the proposed estimators are established in Section 3. Section 4 presents simulation studies to evaluate the empirical performance of the proposed method. Section 5 applies the methodology to the ACTG 175 dataset. Section 6 concludes. Proofs and additional numerical results are provided in the Supplementary Material.

2. Methodology

Let T, C, and X denote the failure (event), censoring, and observed times, respectively. For the ith subject, T_i may not be observed and is subject to right censoring. Throughout the paper, we assume observations from different individuals are independent, but for each individual, its failure time T_i and censoring time C_i are dependent. For right-censored data, we observe $X = \min(T, C)$ and the failure indicator $\delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Suppose that there exists a p-dimensional vector of covariates denoted by $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})^{\mathrm{T}}$, $i = 1, \ldots, n$. Assume that $\{(X_i, \delta_i, \boldsymbol{Z}_i^{\mathrm{T}}) : i = 1, \ldots, n\}$ are independent and identically distributed.

2.1 Copula-based Penalized Likelihood Function

We consider semiparametric PH models to formulate the hazard functions of the event time and the dependent censoring time as follows:

$$\begin{aligned} h_T(t|\boldsymbol{Z}_i) &= h_{0T}(t)\exp(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}), \\ h_C(t|\boldsymbol{Z}_i) &= h_{0C}(t)\exp(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\phi}), \end{aligned} \tag{2.1}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are unknown regression coefficient vectors, and $h_{0C}(\cdot)$ and $h_{0C}(\cdot)$ are unknown baseline hazard functions. Denote their cumulative baseline hazard function by $H_{0T}(\cdot)$ and $H_{0C}(\cdot)$. Roughly speaking, a copula is a function to link two random variables by specifying their dependence structure. In this section, we explore the way in which a copula function (Xu et al., 2018; Chen, 2010) to model the dependence between T and C as follows. Let $K(a, b; \alpha)$ be a copula function with the degree of association parameter α, where $a, b \in [0, 1]$. Note that α can be converted to Kendall's rank correlation coefficient τ between a and b. At the time t, the joint survival function of the event time T and the censoring time C is modeled by the following copula function $K(\cdot)$:

$$S_{T,C}(x,x) = \Pr\left\{T > x, C > x\right\} = K\left(S_T(x), S_C(x); \alpha\right),$$

where $S_T(\cdot)$ and $S_C(\cdot)$ are the marginal survival functions of T and C, respectively.

In what follows, we take the Archimedean copulas as an example to describe the methodology. An extension to other copulas is straightforward. The Archimedean copulas adopt the following functional form: $K(u,v;\alpha) = \phi^{-1}(\phi(u;\alpha) + \phi(v;\alpha))$, where ϕ is called the generator of $K(\cdot)$. It requires that $\phi(u)$ satisfies: (i) $\phi(1) = 0$ and (ii) $\phi(\cdot)$ is a convex and decreasing function with its domain $[0,1]$ and range $[0,\infty]$. Here are some examples of commonly used Archimedean copulas. For instance, the generator of the Frank copula is $\phi(u) = \log\left\{\left(\exp\left(\alpha u\right) - 1\right)/\left(\exp(\alpha) - 1\right)\right\}$, where $-\infty < \alpha < \infty$ and the corresponding Kendall's $\tau = 1 - 4\left(D_1(-\alpha) - 1\right)/\alpha$ with $D_1(\alpha) = \int_0^\alpha t/\left(\exp(t) - 1\right)dt/\alpha$. Then, we have

$$K(u,v;\alpha) = \alpha^{-1}\log\left\{1 + \frac{\left(\exp\left(\alpha u\right) - 1\right)\left(\exp\left(\alpha v\right) - 1\right)}{\exp\left(\alpha\right) - 1}\right\}, \ \alpha \in \mathbb{R}\backslash\{0\}. \quad (2.2)$$

Let $K_1(u,v) = \partial K(u,v)/\partial u$ and $K_2(u,v) = \partial K(u,v)/\partial v$. The likelihood function associated with the PH models in (2.1) is given by

$$L=\prod_{i=1}^{n}\left\{f_T\left(x_i\right)K_1\left(S_T\left(x_i\right),S_C\left(x_i\right)\right)\right\}^{\delta_i}\left\{f_C\left(x_i\right)K_2\left(S_T\left(x_i\right),S_C\left(x_i\right)\right)\right\}^{1-\delta_i}.$$

Denote

$$\ell_{iT}=\log h_{0T}(x_i)+\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}-H_T(x_i)+\log K_1(\exp\left(-H_T\left(x_i\right)\right),\exp\left(-H_C\left(x_i\right)\right)),$$
$$\ell_{iC}=\log h_{0C}(x_i)+\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\phi}-H_C(x_i)+\log K_2(\exp\left(-H_T\left(x_i\right)\right),\exp\left(-H_C\left(x_i\right)\right)),$$

where $H_T(x_i)=H_{0T}(x_i)\exp(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $H_C(x_i)=H_{0C}(x_i)\exp(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\phi})$. Then, the log-likelihood function is

$$\ell=\sum_{i=1}^{n}\left\{\delta_i\ell_{iT}+(1-\delta_i)\ell_{iC}\right\}. \tag{2.3}$$

Manipulating the function $h_{0T}(x)$ or $h_{0C}(x)$ by directly maximizing the log-likelihood is ill-conditioned as it is an infinite-dimensional estimation problem from only a finite number of observations (Xu et al., 2018).

To circumvent this problem, we first approximate $h_{0T}(x)$ and $h_{0C}(x)$ in (2.1) by a function in a finite-dimensional space. Let $\psi_1(x),\ldots,\psi_m(x)$ be the basis functions of this space. Then, we represent $h_{0T}(x)$ and $h_{0C}(x)$ by $h_{0T}(x)=\sum_{u=1}^{m}\theta_u\psi_u(x)$ and $h_{0C}(x)=\sum_{u=1}^{m}\gamma_u\psi_u(x)$, where $\theta_u\geq 0$ and $\gamma_u\geq 0$ for all u. Examples of basis functions include spline, kernel, and indicator functions. Given the complex data structure and the introduction of copula functions, using spline or kernel functions would complicate the data likelihood further. Besides, the selection of knots in the spline approach requires additional effort. As Ma, Heritier and Lô (2014) and Xu et al. (2018) suggested, a piecewise constant function involves relatively simple computation, and the multiplicative-iterative (MI) algorithm is available to estimate the coefficients of the baseline hazard function. Therefore, we develop our algorithm using general indicator basis functions in this study. Indicator basis functions provide a piecewise constant (or step) baseline hazard function. Let $t_{(1)} = \min\{X_i : i = 1, \ldots, n\}$ and $t_{(n)} = \max\{X_i : i = 1, \ldots, n\}$. Suppose sets $\{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ form a partition to $\mathcal{D} = \left[t_{(1)}, t_{(n)}\right]$; i.e., $\cup_{u=1}^m \mathcal{S}_u = \mathcal{D}$ and $\mathcal{S}_u \cap \mathcal{S}_v = \varnothing$ if $u \neq v$. Then, the basis function $\psi_u(t) = I\left(t \in \mathcal{S}_u\right)$, where $I(\cdot)$ is an indicator function.

Let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^{\mathrm{T}}, \boldsymbol{\eta}_2^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\eta}_1 = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\phi}^{\mathrm{T}})^{\mathrm{T}}$, and $\boldsymbol{\eta}_2 = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_m)^T$. We aim to develop an estimation procedure where $h_{0T}(x)$ and $h_{0C}(x)$ are assumed smooth, with smoothness imposed through penalty functions. According to Xu et al. (2018), we use the following roughness penalties: $J(\boldsymbol{\theta}) = \int_t h_{0T}''(x)^2 dt = \boldsymbol{\theta}^{\mathrm{T}} R \boldsymbol{\theta}$ and $J(\boldsymbol{\gamma}) = \int_t h_{0C}''(x)^2 dt = \boldsymbol{\gamma}^{\mathrm{T}} R \boldsymbol{\gamma}$, where R is an $m \times m$ matrix with its (u, v)th element given by $\int_t \psi_u''(x) \psi_v''(x) dt$. Notably, corresponding to discretization, any derivative operation in piecewise constant penalty function should be replaced by difference (Ma, Heritier and Lô, 2014). The square of the first order difference penalty $J(\theta) = \sum_{i=2}^{m}(\theta_i - \theta_{i-1})^2$ is used for all the numerical studies. When many covariates are present, variable selection becomes critical in avoiding the curse of dimensionality, reducing overfitting, and improving model interpretation. Penalized log-likelihood approaches are well-known solutions to the problem for their multiple appealing features. The penalized log-likelihood which we wish to maximize for estimating $\boldsymbol{\eta}$ is

$$\Phi(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \ell(\boldsymbol{\eta}) - \sum_{j=1}^{2p} p_\lambda(|\eta_{1j}|) - J_{h_1,h_2}(\boldsymbol{\eta}_2), \qquad (2.4)$$

where the log-likelihood $\ell(\boldsymbol{\eta})$ is given in (2.3), $p_\lambda(\xi)(\xi > 0)$ is a penalty function that depends on the regularization parameter $\lambda \geq 0$, and roughness penalties $J_{h_1,h_2}(\boldsymbol{\eta}_2) = h_1 J(\boldsymbol{\theta}) + h_2 J(\boldsymbol{\gamma})$, $h_1 \geq 0$ and $h_2 \geq 0$ are the smoothing parameters.

2.2 Variable Selection

Through employing various penalty functions, Equation (2.4) includes many popular variable selection methods. Commonly used penalty functions include LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006), SCAD (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), and MCP (Zhang, 2010), and references therein. Among these penalties, the L_0 penalty is preferred as it possesses desirable theoretical properties and a strong intuitive appeal to penalize the cardinality of a model directly and seeks the most parsimonious model explaining the data (Su et al., 2016; Lv and Fan, 2009; Dicker, Huang and Lin, 2013; Zhao et al., 2020). For instance, the best subset selection (BSS) amounts to using the L_0 penalty of the form: $p_\lambda(|\eta_{1j}|, \lambda) = \lambda I\{\eta_{1j} \neq 0\}$. The BSS often solves

$$\underset{\boldsymbol{\eta_1} \in \mathbb{R}^{2p}}{\arg\min} \left\{-2\ell\left(\boldsymbol{\eta_1}\right) + \lambda_0 \|\boldsymbol{\eta_1}\|_0\right\}, \tag{2.5}$$

where $\|\boldsymbol{\eta}_1\|_0 = \sum_{j=1}^{2p} I(\eta_{1j} \neq 0)$, and a model selection criterion such as AIC or BIC is required to compare models of all choices (Akaike, 1974; Schwarz, 1978). Volinsky and Raftery (2000) suggested replacing the total sample size n with the number of uncensored cases n_0 in the BIC penalty term for censored data because it corresponds to a more realistic prior on the parameter space in the presence of censoring. Therefore, we adopt BIC and set the penalty parameter as $\lambda_0 = \ln(n_0)$. Our numerical studies in Section 4 show that this modified BIC performs satisfactorily. However, the associated variable selection and estimation procedure is nonconvex, and the solution to the corresponding L_0-penalized nonconvex optimization problem involves exhaustive combinatorial best subset search and hence is computationally infeasible for high-dimensional data. These issues render computation a formidable task and implementation challenging in statistical practice.

We consider using the MIC method to tackle this issue. MIC makes sparse estimation through approximating BIC, which essentially involves the approximation of the L_0 norm in its penalty with a hyperbolic tangent function given by $\tanh\left(a|\eta_{1j}|^r\right) = \left\{\exp\left(2a|\eta_{1j}|^r\right) - 1\right\} / \left\{\exp\left(2a|\eta_{1j}|^r\right) + 1\right\}$, where $a > 0$ is a scale parameter that controls the sharpness of the approximation and $r \in \mathbb{N}$ typically takes values of 1 and 2. According to Su et al. (2016), the setting of $r = 1$ leads to a non-smooth optimization problem. Therefore, we choose $r = 2$ to ensure smoothness. Then, we have $\tanh\left(a\eta_{1j}^2\right) = \left\{\exp\left(2a\eta_{1j}^2\right) - 1\right\} / \left\{\exp\left(2a\eta_{1j}^2\right) + 1\right\}$. This step changes the BBS process from discrete to continuous through a continuous smooth surrogate function. Simulation studies conducted by existing works (Su et al., 2016; Han et al., 2019) have demonstrated its satisfactory performance, especially in dealing with complex models.

As mentioned before, the hyperbolic tangent function can be essentially viewed as a continuous approximation of $I(\eta_{1j} \neq 0)$:

$$\lim_{a\to\infty} \frac{\exp\left(2a\eta_{1j}^2\right)-1}{\exp\left(2a\eta_{1j}^2\right)+1} = \begin{cases} 1 & \text{if } \eta_{1j} \neq 0, \\ 0 & \text{if } \eta_{1j} = 0. \end{cases}$$

Besides, the tanh $\left(a\eta_{1j}^2\right)$ is a unit dent function, and the detailed description of the unit dent function can be found in Su et al. (2016). Hence, we denote the MIC penalty as $p_{mic}(|\eta_{1j}|) = \lambda_0 \left\{\exp\left(2a|\eta_{1j}|^2\right)-1\right\} / \left\{\exp\left(2a|\eta_{1j}|^2\right)+1\right\}$. For the sake of revealing the secret of seamless approximation in a nutshell, we have

$$\lambda_0 \|\boldsymbol{\eta}_1\|_0 = \lim_{a\to\infty} \sum_{j=1}^{2p} p_{mic}(|\eta_{1j}|). \tag{2.6}$$

From Equation (2.6), the $p_{mic}(|\beta_j|)$ is a seamless approximation to the L_0 penalty, as $a \to \infty$. For illustration, we plot the MIC penalty in Figure 1. By adjusting a, both unbiasedness and continuity can be easily satisfied. Furthermore, the MIC penalty goes even further, compared to the SELO penalty $p_{selo}(|\beta_j|;\lambda) = \lambda/\log(2)\log\left\{|\beta_j|/(|\beta_j|+\tau)+1\right\}$ with $\tau > 0$ and the regularization parameter $\lambda \geq 0$. We will discuss selecting the tuning parameter a for the MIC procedure in the Supplementary Material. Notably, there is no explicit sparsity in the estimated parameters. Therefore, we set a specific threshold (e.g., 0.0001) to force close-to-zero estimates to precisely zeros to determine the sparsity in the estimated parameters.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

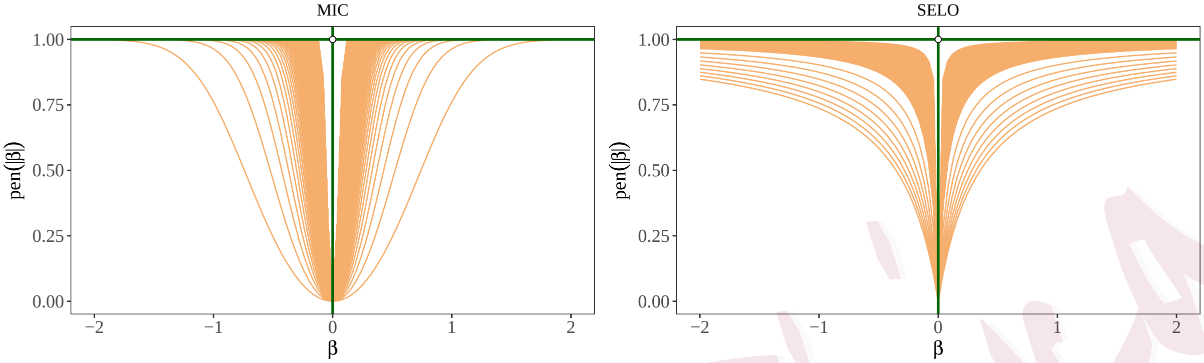

Figure 1: Left: MIC, with the a taking the various values between 1 and 200. Right: SELO, with $\lambda = 1$ and τ taking the values between 0.01 and 0.05.

2.3 Maximum Penalized Likelihood Estimation Procedure

We first separate regression and baseline hazard parameters to simplify future discussions. By applying the MIC method, the penalized log-likelihood function (2.4) can be rewritten as $\Phi(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = 2\ell(\boldsymbol{\eta}) - \lambda_0 \sum_{j=1}^{2p} \tanh(a\eta_{1j}^2) - J_{h_1,h_2}(\boldsymbol{\eta}_2)$. Thus, we want to solve the following constrained optimization problem:

$$(\widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2) = \underset{\boldsymbol{\eta}_1 \in \mathbb{R}^{2p}, \boldsymbol{\eta}_2 \geqslant 0}{\arg\max} \; \Phi\left(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2\right), \tag{2.7}$$

where the inequality is interpreted element-wisely. The Karush-Kuhn-Tucker (KKT) necessary conditions for this constrained optimization are

$$\frac{\partial \Phi}{\partial \beta_j} = 0 \quad \text{and} \quad \frac{\partial \Phi}{\partial \phi_j} = 0, \tag{2.8}$$

$$\begin{aligned} &\frac{\partial \Phi}{\partial \theta_u} = 0 \quad \text{if} \quad \theta_u > 0 \quad \text{and} \quad \frac{\partial \Phi}{\partial \theta_u} < 0 \quad \text{if} \quad \theta_u = 0, \\ &\frac{\partial \Phi}{\partial \gamma_u} = 0 \quad \text{if} \quad \gamma_u > 0 \quad \text{and} \quad \frac{\partial \Phi}{\partial \gamma_u} < 0 \quad \text{if} \quad \gamma_u = 0, \end{aligned} \tag{2.9}$$

for $j = 1, \ldots, p$ and $u = 1, \ldots, m$. To simplify notations, we let $S_{iT} = S_T(x_i), S_{iC} = S_C(x_i), H_{iT} = H_T(x_i)$, and $H_{iC} = H_C(x_i)$. In (2.8), the elements of $\partial\Phi/\partial\boldsymbol{\eta}_1$ are

$$\begin{aligned} \frac{\partial \Phi}{\partial \beta_j} &= \sum_{i=1}^{n} (\delta_i - \delta_i H_{iT} - \Lambda_{i1} S_{iT} H_{iT}) Z_{ij} = 0, \\ \frac{\partial \Phi}{\partial \phi_j} &= \sum_{i=1}^{n} \{(1-\delta_i) - (1-\delta_i) H_{iC} - \Lambda_{i2} S_{iC} H_{iC}\} Z_{ij} = 0, \end{aligned}$$

where

$$\begin{aligned} \Lambda_{i1} &= \delta_i \frac{K_{11}(S_{iT}, S_{iC})}{K_1(S_{iT}, S_{iC})} + (1-\delta_i) \frac{K_{21}(S_{iT}, S_{iC})}{K_2(S_{iT}, S_{iC})}, \\ \Lambda_{i2} &= \delta_i \frac{K_{12}(S_{iT}, S_{iC})}{K_1(S_{iT}, S_{iC})} + (1-\delta_i) \frac{K_{22}(S_{iT}, S_{iC})}{K_2(S_{iT}, S_{iC})}. \end{aligned} \tag{2.10}$$

In above expressions, $\Psi_{iu} = \int_0^{x_i} \psi_u(v) dv$, $S_{iT} = \exp\{-\sum_{u=1}^m \theta_u \Psi_{iu} \exp(Z_i^{\mathrm{T}} \boldsymbol{\beta})\}$, $S_{iC} = \exp\{-\sum_{u=1}^m \theta_u \Psi_{iu} \exp(Z_i^{\mathrm{T}} \boldsymbol{\phi})\}$, $K_{12}(a,b) = \partial^2 K/\partial a \partial b$, $K_{21}(a,b) = \partial^2 K/\partial b \partial a$, $K_{11}(a,b) = \partial^2 K/\partial a^2$ and $K_{22}(a,b) = \partial^2 K/\partial b^2$. Clearly, the elements of $\partial\Phi/\partial\boldsymbol{\eta}_2$ are

$$\frac{\partial\Phi}{\partial\theta_u} = \sum_{i=1}^{n}\left\{\delta_i\frac{\psi_{iu}}{h_{i0T}} - \left(\delta_i + \Lambda_{i1}S_{iT}\right)\Psi_{iu}\exp\left(Z_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right\} - h_1\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_u},$$
$$\frac{\partial\Phi}{\partial\gamma_u} = \sum_{i=1}^{n}\left\{(1-\delta_i)\frac{\psi_{iu}}{h_{i0C}} - \left(1-\delta_i + \Lambda_{i2}S_{iC}\right)\Psi_{iu}\exp\left(Z_i^{\mathrm{T}}\boldsymbol{\phi}\right)\right\} - h_2\frac{\partial J(\boldsymbol{\gamma})}{\partial\gamma_u},$$

where Λ_{i1} and Λ_{i2} are defined in Equation (2.10). An efficient and stable algorithm is a key factor for the successful implementation of the MPL estimation under dependent censoring (Xu et al., 2018). We propose an efficient algorithm to solve Equations (2.8) and (2.9) in the following section.

2.4 A Modified Iterative Algorithm

We use an alternating algorithm similar to Ma (2010) and Xu et al. (2018) to solve Equations (2.8) and (2.9). We call this algorithm the modified Newton multiplicative-iterative (MI) algorithm. Let $\boldsymbol{\eta}_1^{(k)}$ and $\boldsymbol{\eta}_2^{(k)}$ be the estimates of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ at iteration k, respectively. Then, iteration $k+1$ comprises two steps:

- Step 1: With $\boldsymbol{\eta}_2$ setting at $\boldsymbol{\eta}_2^{(k)}$, we update $\boldsymbol{\eta}_1$ using a modified Newton-Raphson algorithm.

- Step 2: With $\boldsymbol{\eta}_1$ fixed at $\boldsymbol{\eta}_1^{(k+1)}$, we update $\boldsymbol{\eta}_2$ using the multiplicative iterative (MI) algorithm (Ma, 2010), where a line search step is included into this iteration to guarantee $\Phi(\boldsymbol{\eta}_1^{(k+1)}, \boldsymbol{\eta}_2)$ increases when moving from $\boldsymbol{\eta}_2^{(k)}$ to $\boldsymbol{\eta}_2^{(k+1)}$. This step also guarantees that each updated $\boldsymbol{\eta}_2$ value respects the non-negativity constraint.

According to (2.7), we first update $\boldsymbol{\eta}_1$ in Step 1 with $\boldsymbol{\eta}_2$ fixed at its current estimate $\boldsymbol{\eta}_2^{(k)}$ from the following regularization problem:

$$\widehat{\boldsymbol{\eta}}_1 = \arg\max_{\boldsymbol{\eta}_1 \in \mathbb{R}^{2p}} 2\ell(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2^{(k)}) - \lambda_0 \sum_{j=1}^{2p} \tanh(a|\eta_{1j}|^2) - J_{h_1,h_2}(\boldsymbol{\eta}_2^{(k)}). \qquad (2.11)$$

To simplify notations, let $\ell(\boldsymbol{\eta}_1) = -2\ell(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2^{(k)})$. Moreover, through some algebraic manipulation, an approximate solution to (2.11) can be equally obtained by solving the following regularization problem:

$$\arg\min_{\boldsymbol{\eta}_1 \in \mathbb{R}^{2p}} \ell\left(\boldsymbol{\eta}_1\right) + \lambda_0 \sum_{j=1}^{2p} \tanh(a|\eta_{1j}|^2), \qquad (2.12)$$

where $\boldsymbol{\eta}_1 = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\phi}^{\mathrm{T}})^{\mathrm{T}}$, and $\lambda_0 = \ln(n_0)$. Furthermore, we assume that $a = O_p(n)$. For ease of notations, we set $\tanh(|\cdot|) = \tanh(a|\cdot|^2)$.

To facilitate a computationally efficient selection procedure, we follow the local quadratic approximation (LQA) approach (Fan and Li, 2001) to approximate the penalty function $\tanh(|\cdot|)$ with $\tanh(|\eta_{1j}|) \approx \tanh(|\alpha_j|) +$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

$\tanh'(|\alpha_j|)/2|\alpha_j|(\eta_j^2 - \alpha_j^2)$, for $\eta_{1j} \approx \alpha_j$. Then we have

$$\sum_{j=1}^{p} p_{mic}(|\eta_{1j}|) \approx \lambda_0 \sum_{j=1}^{2p} \frac{\tanh'(|\alpha_j|)}{2|\alpha_j|}\eta_{1j}^2 + \lambda_0 \sum_{j=1}^{2p} \left(\tanh(|\alpha_j|) - \frac{\tanh'(|\alpha_j|)}{2|\alpha_j|}\alpha_j^2 \right). \tag{2.13}$$

The proposed algorithm starts with a quadratic approximation of $\ell(\boldsymbol{\eta}_1)$ at a generic $\boldsymbol{\alpha}$ by

$$G_t(\boldsymbol{\eta}_1|\boldsymbol{\alpha}) = \ell(\boldsymbol{\alpha}) + (\boldsymbol{\eta}_1 - \boldsymbol{\alpha})^{\mathrm{T}}\ell'(\boldsymbol{\alpha}) + \frac{t}{2}\,\|\boldsymbol{\eta}_1 - \boldsymbol{\alpha}\|_2^2 \tag{2.14}$$

for some pre-specified scaling parameter $t > 0$, where $\|.\|_2$ denotes the L_2 norm and $\ell'(\boldsymbol{\alpha}) = \partial\ell(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$, $\ell''(\boldsymbol{\alpha}) = \partial^2\ell(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}$. The additivity of $G_t(\boldsymbol{\eta}_1|\boldsymbol{\alpha})$ in the components of $\boldsymbol{\eta}_1$ enables us to have a closed-form solution for the minimization problem in (2.12), making the minimization of $G_t(\boldsymbol{\eta}_1|\boldsymbol{\alpha})$ over $\boldsymbol{\eta}_1$ computationally highly efficient.

For $k = 0, 1, 2, \ldots$, set the current estimate $\boldsymbol{\eta}_1^{(k)}$ at iteration k. According to (2.13) and (2.14), with the approximation of the penalty function, an approximate solution to (2.12) can be equally obtained using the following iterative procedure:

$$\arg\min_{\boldsymbol{\eta}_1 \in \mathbb{R}^{2p}} G_t(\boldsymbol{\eta}_1|\boldsymbol{\eta}_1^{(k)}) + \lambda_0 \sum_{j=1}^{2p} \frac{\tanh'(|\eta_{1j}^{(k)}|)}{2|\eta_{1j}^{(k)}|}\eta_{1j}^2. \tag{2.15}$$

Stop the iterations if the sequence of $\{\boldsymbol{\eta}_1^{(k)}\}$ converges. Moreover, through some algebraic manipulation, the iteration (2.15) can be rewritten as

$$\underset{\boldsymbol{\eta}_1\in\mathbb{R}^{2p}}{\operatorname{argmin}}\frac{t}{2}\left\|\boldsymbol{\eta}_1-(\boldsymbol{\eta}_1^{(k)}-t^{-1}\ell'(\boldsymbol{\eta}_1^{(k)}))\right\|_2^2+\lambda_0\sum_{j=1}^{2p}\frac{4a\exp\left(2a|\eta_{1j}^{(k)}|^2\right)}{(\exp\left(2a|\eta_{1j}^{(k)}|^2\right)+1)^2}\eta_{1j}^2. \quad (2.16)$$

The step size t in (2.16) can be determined by the linear search criteria assuring $\ell(\boldsymbol{\eta}_1^{(k+1)}) \leq \ell(\boldsymbol{\eta}_1^{(k)}) - (t\gamma/2)\|\boldsymbol{\eta}_1^{(k+1)} - \boldsymbol{\eta}_1^{(k)}\|_2^2$, where $\gamma \in (0,1)$. One can obtain the MIC estimator $\hat{\boldsymbol{\eta}}_1$ by solving (2.16). The resulting iterative algorithm (Algorithm 1) is presented in the Supplementary Material. The $\hat{\boldsymbol{\eta}}_1$ computed by Algorithm 1 is sparse and updated satisfying $\ell(\hat{\boldsymbol{\eta}}_1) \leq \ell(\boldsymbol{\eta}_1^{(k-1)})$.

It is worth noting that low computational costs are always desirable for regularized variable selection. Unfortunately, most existing variable selection algorithms involve dealing with the singularity of the Hessian matrix (i.e., $[\ell''(\boldsymbol{\eta}_1)]^{-1}$). By contrast, the proposed algorithm is computationally efficient since it can avoid computing the inverse of the Hessian matrix and save computational costs in case of large-scale matrix inversion.

In Step 2, the MI algorithm for $\boldsymbol{\eta}_2$ can be written as

$$\boldsymbol{\eta}_2^{(k+1)} = \boldsymbol{\eta}_2^{(k)} + \omega_2^{(k)}\boldsymbol{S}_2^{(k)}\frac{\partial\Phi(\boldsymbol{\eta}_1^{(k+1)}, \boldsymbol{\eta}_2^{(k)})}{\partial\boldsymbol{\eta}_2}, \quad (2.17)$$

where $\omega_2^{(k)} \in (0,1]$ and $\partial\Phi(\boldsymbol{\eta}_1^{(k+1)}, \boldsymbol{\eta}_2^{(k)})/\partial\boldsymbol{\eta}_2$ is the gradient of $\Phi(\boldsymbol{\eta}_1^{(k+1)}, \boldsymbol{\eta}_2)$ evaluated at $\boldsymbol{\eta}_2^{(k)}$ and $\boldsymbol{S}_2^{(k)}$ is a diagonal matrix given by $\boldsymbol{S}_2^{(k)} = \mathrm{diag}(\boldsymbol{S}_{21}^{(k)}, \boldsymbol{S}_{22}^{(k)})$. Expressions for $\partial\Phi/\partial\boldsymbol{\eta}_2$, $\boldsymbol{S}_{21}^{(k)} = (\theta_1^{(k)}/\xi_{11}^{(k)}, \ldots, \theta_m^{(k)}/\xi_{1m}^{(k)})$ and $\boldsymbol{S}_{22}$ is given as $\boldsymbol{S}_{22}^{(k)} = (\gamma_1^{(k)}/\xi_{21}^{(k)}, \ldots, \gamma_m^{(k)}/\xi_{2m}^{(k)})$. Here,

$$\begin{aligned}
\xi_{1u}^{(k)} &= \sum_{i=1}^{n}\left(\delta_{iT}\Psi_{iu}e^{Z_i^{\mathrm{T}}\boldsymbol{\beta}^{(k+1)}} + \Lambda_{i1+}^{(k)}S_{iT}^{(k)}\Psi_{iu}e^{Z_i^{\mathrm{T}}\boldsymbol{\beta}^{(k+1)}}\right) + h_1\left[\frac{\partial J(\boldsymbol{\theta}^{(k)})}{\partial\theta_u}\right]^{+} + \epsilon, \\
\xi_{2u}^{(k)} &= \sum_{i=1}^{n}\left(\delta_{iC}\Psi_{iu}e^{Z_i^{\mathrm{T}}\boldsymbol{\beta}^{(k+1)}} + \Lambda_{i2+}^{(k)}S_{iC}^{(k)}\Psi_{iu}e^{Z_i^{\mathrm{T}}\boldsymbol{\phi}^{(k+1)}}\right) + h_2\left[\frac{\partial J\left(\boldsymbol{\gamma}^{(k)}\right)}{\partial\gamma_u}\right]^{+} + \epsilon,
\end{aligned} \tag{2.18}$$

where $[c]^+ = \max(c, 0)$, and ϵ is a small non-zero constant (i.e 10^{-5}) used to avoid $\xi_{1u}^{(k)}$ and $\xi_{2u}^{(k)}$ being zero. In Equation (2.18),

$$\begin{aligned}
\Lambda_{i1+} &= \delta_i\frac{[K_{11}(S_{iT}, S_{iC})]^{+}}{K_1(S_{iT}, S_{iC})} + (1-\delta_i)\frac{K_{21}(S_{iT}, S_{iC})}{K_2(S_{iT}, S_{iC})}, \\
\Lambda_{i2+} &= (1-\delta_i)\frac{[K_{22}(S_{iT}, S_{iC})]^{+}}{K_2(S_{iT}, S_{iC})} + \delta_i\frac{K_{12}(S_{iT}, S_{iC})}{K_1(S_{iT}, S_{iC})},
\end{aligned}$$

and they are non-negative since K_{12}, K_{21}, K_2, K_1, and K are all non-negative. It is clear that if $\boldsymbol{\eta}_2^{(k)} \geq 0$, then $\boldsymbol{\eta}_2^{(k+1)}$ given by Equation (2.17) is also non-negative for any $\omega_2^{(k)} \in (0,1]$. The step size $\omega_2^{(k)}$ in (2.17) can be determined by the Armijo rule assuring $\Phi(\boldsymbol{\eta}_1^{(k+1)}, \boldsymbol{\eta}_2^{(k+1)}) \geq \Phi(\boldsymbol{\eta}_1^{(k+1)}, \boldsymbol{\eta}_2^{(k)})$. Ma, Heritier and Lô (2014) showed the convergence properties of the Newton-MI algorithm.

3. Theoretical Results

In this section, we investigate the asymptotic properties of the proposed method. The theoretical results are summarized in Theorems 1 to 3, and the proofs are provided in the Supplementary Material.

Let $\boldsymbol{\pi}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, h^0_{0T}(x), h^0_{0C}(x))$ and $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{h}_{0T}(x), \hat{h}_{0C}(x))$ be the sets for the true parameters and their MPL estimates. Let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\phi}_0 = (\boldsymbol{\phi}_{10}^{\mathrm{T}}, \boldsymbol{\phi}_{20}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{20} = \boldsymbol{\phi}_{20} = 0$. Without loss of generality, we define the active set $\mathscr{C}_1 = \{j : \beta_{j0} \neq 0, 1 \leq j \leq p\}$ and $\mathscr{C}_2 = \{j : \phi_{j0} \neq 0, 1 \leq j \leq p\}$. Then, we have $s_1 = \|\mathscr{C}_1\|_0$ and $s_2 = \|\mathscr{C}_2\|_0$. $\boldsymbol{\beta}_{10}$ consists of all s_1 nonzero elements of $\boldsymbol{\beta}_0$, and $\boldsymbol{\phi}_{10}$ consists of all s_2 nonzero elements of $\boldsymbol{\phi}_0$. For two different $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$, we define the norm $\rho(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \left\{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2\|_2^2 + \|h^1_{0T} - h^2_{0T}\|_2^2 + \|h^1_{0C} - h^2_{0C}\|_2^2\right\}^{1/2}$.

The results in Theorem 1 state that the MPL estimators converge to their true values when the number of bins $m \to \infty$ but slower than $n \to \infty$ in that $m/n \to 0$, and both the regularization parameter $\lambda_n = \lambda_0/n$ and scaled smoothing values $\mu_{1n} = h_1/n, \mu_{2n} = h_2/n$ go to zero when $n \to \infty$.

Theorem 1. *Assume that Assumptions A1 to A6 provided in the Supplementary Material hold, and $h_{0T}(x)$ and $h_{0C}(x)$ have up to $r \geq 1$ derivatives. Assume $m = n^{\nu}$ where $0 < \nu < 1/2$, and λ_n, μ_{1n} and $\mu_{2n} \to 0$ as $n \to \infty$.*

Then, when $n \to \infty$,

(i) $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \to 0$ *(a.s.) and* $\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\|_2 \to 0$ *(a.s.);*

(ii) $\sup_{t\in[t_{(1)},t_{(n)}]} |\hat{h}_{nT}(t) - h^0_{0T}(t)| \to 0$ *(a.s.) and* $\sup_{t\in[t_{(1)},t_{(n)}]} |\hat{h}_{nC}(t) - \lambda^0_{0C}(t)| \to 0$ *(a.s.).*

Theorem 1 guarantees that the MPL estimators converge to their true values under some regularity conditions. Next, we construct the convergence rate for the estimated parameters.

Theorem 2. *(Rate of convergence) Suppose that the assumptions listed in Theorem 1 hold. Then, as $n \to \infty$, we have*

$$\rho(\hat{\boldsymbol{\pi}}_n, \boldsymbol{\pi}_0) = O_p\left(n^{-(1-\nu)/2} + n^{-\zeta\nu}\right), 0 < \zeta \leq 1.$$

In particular, by taking $\nu = 1/(2\zeta+1)$, we have $\rho(\hat{\boldsymbol{\pi}}_n, \boldsymbol{\pi}_0) = O_p\left(n^{-\zeta/(2\zeta+1)}\right)$.

Let $\hat{\boldsymbol{\eta}}$ denote the constrained MPL estimate of $\boldsymbol{\eta}$, where $\boldsymbol{\theta} \geq 0$ and $\boldsymbol{\gamma} \geq 0$. Let $\boldsymbol{\eta}_0$ be the true parameter set with a fixed m. Following Xu et al. (2018), we give the asymptotic properties of the MPL estimators. The matrix $\boldsymbol{U}$ defined in Assumption B5 in the Supplementary Material indicates the active sets and constraints. Note that $\boldsymbol{U}^\top\boldsymbol{U} = I_{[s_1+s_2+2m-q-l]\times[s_1+s_2+2m-q-l]}$, where q and l are the numbers of active constraints from $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, respectively, and s_1 and s_2 are the numbers of nonzero elements of $\boldsymbol{\beta}_0$ and $\boldsymbol{\phi}_0$, respectively. Let $\boldsymbol{I}_0(\boldsymbol{\eta}) = -E_{\boldsymbol{\eta}_0}\left[\partial^2 l(\boldsymbol{\eta})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}\right]$ is the expected information matrix and $\boldsymbol{G}_0(\boldsymbol{\eta}) = \boldsymbol{I}_0(\boldsymbol{\eta}) + \mu_{1n}\partial^2 J_1(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}} + \mu_{2n}\partial^2 J_2(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}} + \lambda\partial^2 p_\lambda(\boldsymbol{\eta_1})/\partial\boldsymbol{\eta_1}\partial\boldsymbol{\eta_1}^{\mathrm{T}}$.

Theorem 3. *Under the conditions of Theorem 1 and Assumptions B1 to B5 provided in the Supplementary Material, both μ_{1n} and μ_{2n} are $o\left(n^{-1/2}\right)$. Suppose $n_0 = O_p(n)$, the MPL estimators $\hat{\boldsymbol{\eta}}$ must satisfy:*

(i) With probability tending to one, $(\hat{\boldsymbol{\beta}}_{20}^{\mathrm{T}}, \hat{\boldsymbol{\phi}}_{20}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{0}$;

(ii) $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)$ converges in distribution to $N\left(\mathbf{0}_{2\times(m+p)}, \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\right)$ when $n \to \infty$, with $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0) = \widetilde{\boldsymbol{G}}_0\left(\boldsymbol{\eta}_0\right)^{-1}\boldsymbol{I}_0\left(\boldsymbol{\eta}_0\right)\left\{\widetilde{\boldsymbol{G}}_0\left(\boldsymbol{\eta}_0\right)^{-1}\right\}^{\mathrm{T}}$ and $\widetilde{\boldsymbol{G}}_0\left(\boldsymbol{\eta}_0\right)^{-1} = \boldsymbol{U}\left\{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{G}_0(\boldsymbol{\eta})\boldsymbol{U}\right\}^{-1}\boldsymbol{U}^{\mathrm{T}}$.

In practice, $\boldsymbol{\eta}_0$ is generally unavailable, and we can replace it with $\hat{\boldsymbol{\eta}}$ due to the strong consistency result. Theorem 3 has practical values since it accommodates nonzero smoothing values and active constraints.

4. Simulation Study

This section presents simulation studies to assess the finite sample performance of the proposed MPL method. For comparison, we also evaluate the penalized partial likelihood (PL) method of Cox (Cox, 1972) and conduct simulations under full (with all predictors) and oracle (with actual predictors) PH models. Since our work is the first variable selection method with

MIC penalty function for PH models in the presence of dependent censoring, it is uncertain how other sparse estimation methods, such as LASSO or SCAD, can be extended to the current model context. For this reason, we do not compare our method with other competitive approaches.

We use the following measures to assess the performance of the selection procedure: (i) The percentage of correct model selection (Pcorr); (ii) The mean weighted squared error (MSE); i.e., $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\Sigma(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, where Σ is the population covariance matrix of the covariates; (iii) The average number of false positives (i.e., the average number of incorrectly included variables), denoted by F_+; (iv) The average number of false negatives (i.e., the average number of incorrectly excluded variables), denoted by F_-; (v) the averaged model size (`Size`). A powerful variable selection procedure must ensure that Pcorr is close to one and F_+, F_-, and MSE are close to zero. Also, the estimated model size (i.e., the estimated number of covariates) should be close to the true model size. The tuning parameter a is selected using an extended BIC procedure described in the Supplementary Material.

We set the sample size $n = 200$ and 400, and the total number of covariates $p = 10$ and 20. All the simulation results given below are based on 500 replications. According to Ma, Heritier and Lô (2014), we use the equal bin event count strategy to define the indicator functions, where the common event count of each bin is denoted by n_0. In practice, we set $n_0 = 2$ for $n = 200$ and $n_0 = 4$ for $n = 400$ and the number of bins $m = n/n_0$. We apply a single tuning parameter to all the replications for the optimal smoothing parameter h_1 and h_2. Based on a replicated sample and a chosen m, the optimal smoothing parameter h_1 and h_2 are estimated through BIC. Then, we obtain the initial regression coefficient estimates in the PH models for failure and dependent censoring times using the MIC approach with $r = 2$ and update them iteratively through the LQA until convergence. We use a threshold of 0.0001 to determine the sparsity of the estimated parameters. In practice, we include this additional step to Algorithm 1 presented in the Supplementary Material.

We conduct relevant simulation experiments listed below. Two different scenarios for the true sparse regression coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\phi}_0$ are considered as follows:

$$\text{Case A: } \boldsymbol{\beta}_0 = (0.5; 0; -0.5; 0; 0; 0.5; \mathbf{0}_{p-6}),\ \boldsymbol{\phi}_0 = (0; 0.5; 0; -0.5; \mathbf{0}_{p-4}).$$

$$\text{Case B: } \boldsymbol{\beta}_0 = (1.0; 0; -1.0; 0; 0; 1.0; \mathbf{0}_{p-6}),\ \boldsymbol{\phi}_0 = (0; 1.0; 0; -1.0; \mathbf{0}_{p-4}).$$

Cases A and B correspond to situations with small and large covariate effects, respectively. The data used in the simulation are obtained as follows. For a given p, we generate covariate Z from a multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$. We assume a slight and moderate correlation between the covariates by taking $\rho = 0.25$ and 0.5, respectively. Our settings involve both dependent and independent censoring. Under the dependent censoring case, the marginal hazards for T and C are set as follows: $h_T(t|\boldsymbol{Z}) = h_{0T}(t)\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}\right)$, $h_C(t|\boldsymbol{Z}) = h_{0C}(t)\exp\left(\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{Z}\right)$, where the baseline hazards are $h_{0C}(t) = 1/5$ and $h_{0T}(t) = 2t/\lambda_t^2$. Here, λ_t can be chosen for a specified censoring rate (CR) of 30%. For each T_i, the independent censoring time C_i is generated from the uniform distribution $U(0, c_0)$ with c_0 chosen to obtain a CR of either 25% or 45%.

Under dependent censoring, Tables 1 and 2 summarize the MSE and variable selection results in Cases A and B, respectively. In addition, Table 3 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 proposed and PL methods in Case A under dependent censoring. Table S1 reports the MSE and variable selection results in Case A under independent censoring.

Table 1 indicates that Pcorr is equal to 100%, F_+ and F_- are very close to zero, and `Size` is sufficiently close to 3 for our method. This result sug-

Table 1: 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, 10)	Full	0.00	0.105	6.992	0.000	9.992	0.00	0.174	6.992	0.000	9.992
		PL	93.80	0.039	0.052	0.012	3.040	85.00	0.083	0.166	0.000	3.166
		MPL	100.00	0.027	0.000	0.000	3.000	100.00	0.054	0.000	0.000	3.000
		Oracle	100.00	0.029	0.000	0.000	3.000	100.00	0.058	0.000	0.000	3.000
	(0.25, 20)	Full	0.00	0.249	16.990	0.000	19.990	0.00	0.348	16.988	0.000	19.988
		PL	87.40	0.051	0.132	0.008	3.124	77.60	0.107	0.268	0.006	3.262
		MPL	100.00	0.031	0.000	0.000	3.000	100.00	0.062	0.000	0.000	3.000
		Oracle	100.00	0.032	0.000	0.000	3.000	100.00	0.063	0.000	0.000	3.000
	(0.5, 10)	Full	0.00	0.105	6.996	0.000	9.996	0.00	0.161	6.994	0.000	9.994
		PL	86.40	0.046	0.124	0.032	3.092	77.80	0.083	0.272	0.004	3.268
		MPL	99.60	0.029	0.000	0.004	2.996	99.80	0.051	0.000	0.002	2.998
		Oracle	100.00	0.030	0.000	0.000	3.000	100.00	0.052	0.000	0.000	3.000
	(0.5, 20)	Full	0.00	0.247	16.990	0.000	19.990	0.00	0.329	16.984	0.000	19.984
		PL	81.80	0.058	0.218	0.030	3.188	70.80	0.111	0.406	0.010	3.396
		MPL	100.00	0.032	0.000	0.000	3.000	100.00	0.058	0.000	0.000	3.000
		Oracle	100.00	0.031	0.000	0.000	3.000	100.00	0.059	0.000	0.000	3.000
$n = 400$	(0.25, 10)	Full	0.00	0.048	6.994	0.000	9.994	0.00	0.102	6.996	0.000	9.996
		PL	99.80	0.016	0.002	0.000	3.002	97.80	0.047	0.022	0.000	3.022
		MPL	100.00	0.016	0.000	0.000	3.000	100.00	0.040	0.000	0.000	3.000
		Oracle	100.00	0.015	0.000	0.000	3.000	100.00	0.040	0.000	0.000	3.000
	(0.25, 20)	Full	0.00	0.100	16.986	0.000	19.986	0.00	0.167	16.962	0.000	19.962
		PL	99.80	0.019	0.002	0.000	3.002	96.60	0.056	0.034	0.000	3.034
		MPL	100.00	0.017	0.000	0.000	3.000	100.00	0.046	0.000	0.000	3.000
		Oracle	100.00	0.016	0.000	0.000	3.000	100.00	0.043	0.000	0.000	3.000
	(0.5, 10)	Full	0.00	0.048	6.994	0.000	9.994	0.00	0.090	7.000	0.000	10.000
		PL	99.40	0.016	0.004	0.002	3.002	96.20	0.043	0.042	0.000	3.042
		MPL	100.00	0.015	0.000	0.000	3.000	100.00	0.036	0.000	0.000	3.000
		Oracle	100.00	0.014	0.000	0.000	3.000	100.00	0.036	0.000	0.000	3.000
	(0.5, 20)	Full	0.00	0.100	16.978	0.000	19.978	0.00	0.153	16.988	0.000	19.988
		PL	99.00	0.018	0.010	0.000	3.010	93.80	0.051	0.062	0.000	3.062
		MPL	100.00	0.017	0.000	0.000	3.000	100.00	0.040	0.000	0.000	3.000
		Oracle	100.00	0.016	0.000	0.000	3.000	100.00	0.039	0.000	0.000	3.000

Table 2: 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
$n = 200$	$(0.25, 10)$	Full	0.00	0.121	6.994	0.000	9.994	0.00	0.203	6.992	0.000	9.992
		PL	92.80	0.045	0.072	0.000	3.072	81.60	0.098	0.198	0.000	3.198
		MPL	100.00	0.023	0.000	0.000	3.000	100.00	0.063	0.000	0.000	3.000
		Oracle	100.00	0.039	0.000	0.000	3.000	100.00	0.069	0.000	0.000	3.000
	$(0.25, 20)$	Full	0.00	0.277	16.988	0.000	19.988	0.00	0.394	16.988	0.000	19.988
		PL	85.80	0.070	0.164	0.000	3.164	74.60	0.144	0.332	0.000	3.332
		MPL	100.00	0.031	0.000	0.000	3.000	100.00	0.088	0.000	0.000	3.000
		Oracle	100.00	0.039	0.000	0.000	3.000	100.00	0.071	0.000	0.000	3.000
	$(0.5, 10)$	Full	0.00	0.119	6.996	0.000	9.996	0.00	0.200	6.996	0.000	9.996
		PL	86.40	0.048	0.152	0.004	3.152	70.40	0.106	0.374	0.000	3.374
		MPL	100.00	0.024	0.000	0.000	3.000	100.00	0.064	0.000	0.000	3.000
		Oracle	100.00	0.035	0.000	0.000	3.000	100.00	0.069	0.000	0.000	3.000
	$(0.5, 20)$	Full	0.00	0.267	16.990	0.000	19.990	0.00	0.376	16.992	0.000	19.992
		PL	78.20	0.073	0.292	0.004	3.292	62.60	0.145	0.510	0.000	3.510
		MPL	100.00	0.028	0.000	0.000	3.000	100.00	0.082	0.000	0.000	3.000
		Oracle	100.00	0.037	0.000	0.000	3.000	100.00	0.071	0.000	0.000	3.000
$n = 400$	$(0.25, 10)$	Full	0.00	0.054	6.990	0.000	9.990	0.00	0.116	6.992	0.000	9.992
		PL	99.80	0.017	0.002	0.000	3.002	95.40	0.048	0.050	0.000	3.050
		MPL	100.00	0.010	0.000	0.000	3.000	100.00	0.043	0.000	0.000	3.000
		Oracle	100.00	0.018	0.000	0.000	3.000	100.00	0.043	0.000	0.000	3.000
	$(0.25, 20)$	Full	0.00	0.109	16.984	0.000	19.984	0.00	0.189	16.988	0.000	19.988
		PL	99.00	0.021	0.004	0.000	3.000	93.80	0.064	0.068	0.000	3.068
		MPL	100.00	0.013	0.000	0.000	3.000	100.00	0.057	0.000	0.000	3.000
		Oracle	100.00	0.018	0.000	0.000	3.000	100.00	0.046	0.000	0.000	3.000
	$(0.5, 10)$	Full	0.00	0.054	6.994	0.000	9.994	0.00	0.115	6.996	0.000	9.996
		PL	99.20	0.016	0.006	0.000	3.006	91.20	0.051	0.094	0.000	3.094
		MPL	100.00	0.011	0.000	0.000	3.000	100.00	0.045	0.000	0.000	3.000
		Oracle	100.00	0.017	0.000	0.000	3.000	100.00	0.044	0.000	0.000	3.000
	$(0.5, 20)$	Full	0.00	0.106	16.986	0.000	19.986	0.00	0.183	16.986	0.000	19.986
		PL	98.80	0.020	0.012	0.004	3.012	89.80	0.063	0.120	0.000	3.120
		MPL	100.00	0.013	0.000	0.000	3.000	100.00	0.054	0.000	0.000	3.000
		Oracle	100.00	0.018	0.000	0.000	3.000	100.00	0.046	0.000	0.000	3.000

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

Table 3: 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
$n = 200$	$(0.25, 10)$	β_1	0.019	0.057	0.061	0.948	0.018	0.098	0.034	0.540
		β_3	-0.025	0.065	0.061	0.914	-0.016	0.102	0.033	0.470
		β_6	0.026	0.060	0.061	0.934	0.016	0.096	0.032	0.466
		ϕ_2	-0.032	0.066	0.085	0.950	- -	- -	- -	- -
		ϕ_4	0.025	0.062	0.085	0.964	- -	- -	- -	- -
	$(0.25, 20)$	β_1	0.019	0.057	0.061	0.948	0.018	0.098	0.034	0.540
		β_3	-0.025	0.065	0.061	0.914	-0.016	0.102	0.033	0.470
		β_6	0.026	0.060	0.061	0.934	0.016	0.096	0.032	0.466
		ϕ_2	-0.032	0.066	0.085	0.950	- -	- -	- -	- -
		ϕ_4	0.025	0.062	0.085	0.964	- -	- -	- -	- -
	$(0.5, 10)$	β_1	0.002	0.057	0.063	0.952	0.002	0.095	0.035	0.482
		β_3	-0.002	0.061	0.063	0.940	-0.012	0.093	0.034	0.500
		β_6	-0.017	0.057	0.061	0.940	-0.014	0.092	0.031	0.422
		ϕ_2	-0.049	0.065	0.086	0.946	- -	- -	- -	- -
		ϕ_4	0.047	0.063	0.085	0.932	- -	- -	- -	- -
	$(0.5, 20)$	β_1	0.009	0.057	0.062	0.958	0.013	0.102	0.034	0.496
		β_3	-0.016	0.064	0.063	0.930	-0.002	0.106	0.032	0.442
		β_6	0.025	0.059	0.061	0.932	0.003	0.104	0.029	0.420
		ϕ_2	-0.032	0.061	0.086	0.960	- -	- -	- -	- -
		ϕ_4	0.030	0.067	0.086	0.960	- -	- -	- -	- -
$n = 400$	$(0.25, 10)$	β_1	-0.004	0.033	0.043	0.978	-0.010	0.059	0.028	0.592
		β_3	-0.001	0.035	0.043	0.970	-0.015	0.063	0.026	0.578
		β_6	0.008	0.038	0.043	0.958	0.009	0.064	0.026	0.544
		ϕ_2	-0.037	0.033	0.057	0.952	- -	- -	- -	- -
		ϕ_4	0.037	0.032	0.057	0.956	- -	- -	- -	- -
	$(0.25, 20)$	β_1	0.005	0.036	0.043	0.968	0.002	0.064	0.027	0.576
		β_3	-0.012	0.037	0.043	0.962	-0.001	0.069	0.026	0.506
		β_6	0.011	0.036	0.043	0.954	-0.003	0.066	0.026	0.544
		ϕ_2	-0.030	0.033	0.057	0.976	- -	- -	- -	- -
		ϕ_4	0.033	0.031	0.057	0.966	- -	- -	- -	- -
	$(0.5, 10)$	β_1	-0.014	0.034	0.044	0.970	-0.015	0.062	0.035	0.576
		β_3	0.006	0.036	0.044	0.966	0.024	0.065	0.028	0.540
		β_6	0.011	0.038	0.043	0.956	-0.014	0.064	0.027	0.480
		ϕ_2	-0.038	0.039	0.060	0.942	- -	- -	- -	- -
		ϕ_4	0.037	0.039	0.060	0.952	- -	- -	- -	- -
	$(0.5, 20)$	β_1	-0.005	0.036	0.044	0.970	-0.004	0.064	0.029	0.568
		β_3	-0.006	0.038	0.044	0.956	0.010	0.069	0.026	0.512
		β_6	0.012	0.038	0.043	0.946	-0.010	0.066	0.026	0.506
		ϕ_2	-0.031	0.039	0.060	0.960	- -	- -	- -	- -
		ϕ_4	0.033	0.039	0.060	0.948	- -	- -	- -	- -

gests that our method can accurately discover the sparse representation of Models (2.1). Moreover, our method 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 discrepancy increases as the correlation between the survival and censoring times increases from 0.2 to 0.5. In Case B, important variables have larger effects than in Case A. Table 2 shows similar results to Table 1. In addition, from Tables 1 and 2, 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 10 to 20.

Table 3 presents the results of the nonzero coefficients in Case A. We use the standard error formula in Zhang and Lu (2007) to estimate the standard errors for the PL estimates. The proposed MPL keeps a better agreement between SE and ASE values than the PL method, and the coverage probabilities of the 95% confidence intervals yielded by our method are closer to the nominal level than the PL method. For all methods, the estimated and sample standard errors decrease as n increases from 200 to 400.

Table S1 in the Supplementary Material summarizes MSE and variable selection results in Case A under independent censoring. As expected, the performance of MPL and PL is comparable when censoring is independent, and they both perform better as n increases. When $n = 400$, even as the model size or CR increases, both methods still have high accuracy in estimation and variable selection.

We also consider other settings with larger p, selecting only $\boldsymbol{\beta}$ and using a spline baseline hazard. Tables S2–S6 in the Supplementary Material suggest that our method performs well when p is relatively large and when using a different baseline hazard but performs better when selecting $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ simultaneously.

The code for implementing the simulation study is publicly available at https://github.com/zili-liu/mpl_informative_censoring.

5. The ACTG 175 study

We applied the proposed method to the ACTG 175 study, which contains 2,467 HIV-1-infected patients with CD4 counts between 200 and 500 per cubic millimeter (Hammer et al., 1996). The ACTG 175 study evaluated treatment with either a single nucleoside or two nucleosides, where patients were randomized to one of four daily antiretroviral regimens in equal proportions: zidovudine only, zidovudine plus didanosine, zidovudine plus

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

zalcitabine, and didanosine only (Hammer et al., 1996). Enrolment began in December 1991 and was closed in October 1992. Patients were scheduled to be followed until November 1994. As a result, all subjects were scheduled for at least two years of follow-up. The CD4 counts were collected at baseline, week 8, and then every 12 weeks thereafter. The primary endpoint was time to 50% decline in CD4 from baseline, as confirmed by a second CD4 count within 3 to 21 days, AIDS or death. The censoring rate of the dataset is approximately 76%.

Possible risk factors include those relevant to patients' characteristics, which consist of four continuous covariates: weight (kg), Karnofsky score (scale of 0–100), CD4 count (cells/mm^3) at baseline and CD8 count (cells/mm^3) at baseline, and ten binary covariates: age > 50 ($1 =$ yes), hemophilia ($1 =$ yes), homosexual activity ($1 =$ yes), non-zidovudine antiretroviral therapy prior to initiation of study treatment ($1 =$ yes), race ($1 =$ non-white), gender ($1 =$ male), antiretroviral history ($1 =$ experienced), symptomatic status ($1 =$ symptomatic), treatment indicator ($0 =$ zidovudine only, $1 =$ others), and indicator of off-treatment before 96 ± 5 weeks ($1 =$ yes). For each subject, T was defined as the time to the primary endpoint that would be observed under full compliance, and the censoring time, C, was defined as the shorter time to loss to follow-up and time to discontinuation of assigned therapy. Scharfstein and Robins (2002) showed that those reporting injection-drug use and with lower CD4 cell counts, lower Karnofsky scores, and symptoms of HIV infection at enrolment were significantly more likely to discontinue treatment before the study ended, suggesting informative censoring. Therefore, unlike Huang and Zhang (2008) who selected covariates using a stepwise selection algorithm under independent censoring, we conducted a simultaneous estimation and variable selection in the presence of informative censoring.

We analyzed the ACTG 175 data using the proposed MPL and competing PL methods. For MPL method, we analyzed the data by assuming a Frank copula model for dependent censoring and selected the correlation parameter τ using BIC. That is, we fitted the model repeatedly using Kendall's $\tau = 0$ (i.e., independent censoring), 0.2 (weak dependence between event and censoring times), 0.5 (moderate dependence between event and censoring times), and 0.7 (strong dependence between event and censoring times). Then, we chose the optimal τ using BIC. In the analysis, we adopted an equal number of observations in each bin to determine the bins and construct a piecewise constant function to approximate the baseline hazard. For the optimal smoothing parameter h_1 and h_2, given a chosen m, the optimal smoothing parameter h_1 and h_2 were determined through

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

BIC. The initial estimates of the regression coefficient of the PH models were obtained using the MIC approach.

Table 4: Variable selection and estimation results for the ACTG 175 data. Abbreviations: SE, standard error; CI, 95% confidence interval.

	MPL				PL			
	$\hat{\boldsymbol{\beta}}$	SE	CI	p-value	$\hat{\boldsymbol{\beta}}$	SE	CI	p-value
age	0.3297	0.1367	(0.0617, 0.5977)	0.0159	– –	– –	– –	– –
wtkg	– –	– –	– –	– –	– –	– –	– –	– –
hemo	– –	– –	– –	– –	– –	– –	– –	– –
homo	0.1566	0.0560	(0.0470, 0.2663)	0.0051	– –	– –	– –	– –
karnof	-0.1394	0.0284	(-0.1950, -0.0837)	< 0.0001	– –	– –	– –	– –
oprior	– –	– –	– –	– –	– –	– –	– –	– –
race	– –	– –	– –	– –	– –	– –	– –	– –
gender	– –	– –	– –	– –	– –	– –	– –	– –
str2	0.3424	0.0582	(0.2283, 0.4564)	< 0.0001	0.3376	0.0564	(0.2269, 0.4482)	< 0.0001
symptom	0.3185	0.0573	(0.2062, 0.4309)	< 0.0001	0.3096	0.0816	(0.1496, 0.4695)	< 0.0001
treat	-0.6336	0.0554	(-0.7421, -0.5251)	< 0.0001	-0.5109	0.0631	(-0.6345, -0.3872)	< 0.0001
offtrt	0.6085	0.0618	(0.4875, 0.7296)	< 0.0001	0.5056	0.0234	(0.4598, 0.5515)	< 0.0001
cd40	-0.4443	0.0405	(-0.5236, -0.3650)	< 0.0001	-0.4123	0.0376	(-0.4860, -0.3386)	< 0.0001
cd80	0.2341	0.0296	(0.1761, 0.2921)	< 0.0001	– –	– –	– –	– –

Table 4 shows that the MPL method selected nine variables: age, homo, karnof, str2, symptom, treat, offtrt, cd40, and cd80, in which five variables, including age, str2, symptom, treat, and cd40, are similar to those selected in Huang and Zhang (2008). Based on Table 4, we have several observations. First, age at enrolment has a significantly positive effect on the hazards of $\geq$ 50 percent decline in the CD4 cell count, indicating that older patients have higher risks of developing AIDS. Second, homosexual activity, symptomatic HIV/AIDS, history of antiretroviral infection, off-treatment before

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

96 ± 5 weeks, and CD8 cell count also significantly increase AIDS hazards. Patients reporting antiretroviral infection, with previous homosexual activity, symptoms of HIV infection at enrolment, or high levels of CD8 cell count, and discontinuing the treatment before 96 ± 5 weeks are more predisposed to AIDS. Third, sex, race, weight, hemophilia, and previous non-zidovudine antiretroviral therapy do not exert a significant effect on AIDS development. In contrast, karnof and cd40 have significantly adverse effects on AIDS hazards, implying that poor health and low levels of the CD4 cell count increase the risk of AIDS. Besides, the treatment indicator also has a significantly negative impact on developing AIDS hazards. That is, regimens combining zidovudine with other antiretroviral agents reduces AIDS hazards. In comparison, the PL method only selected five variables: str2, symptom, treat, offer, and cd40, and failed to identify age, homo, karnof, and cd80 as the risk factors of AIDS. Therefore, the proposed MPL procedure can identify essential predictors more efficiently than the PL method because it accounts for a range of possible Kendall's τ values in the presence of informative censoring. These findings have public health implications, especially for the aggressive control of risk factors to prevent AIDS or other complications for HIV-1-infected patients and to improve their quality of life.

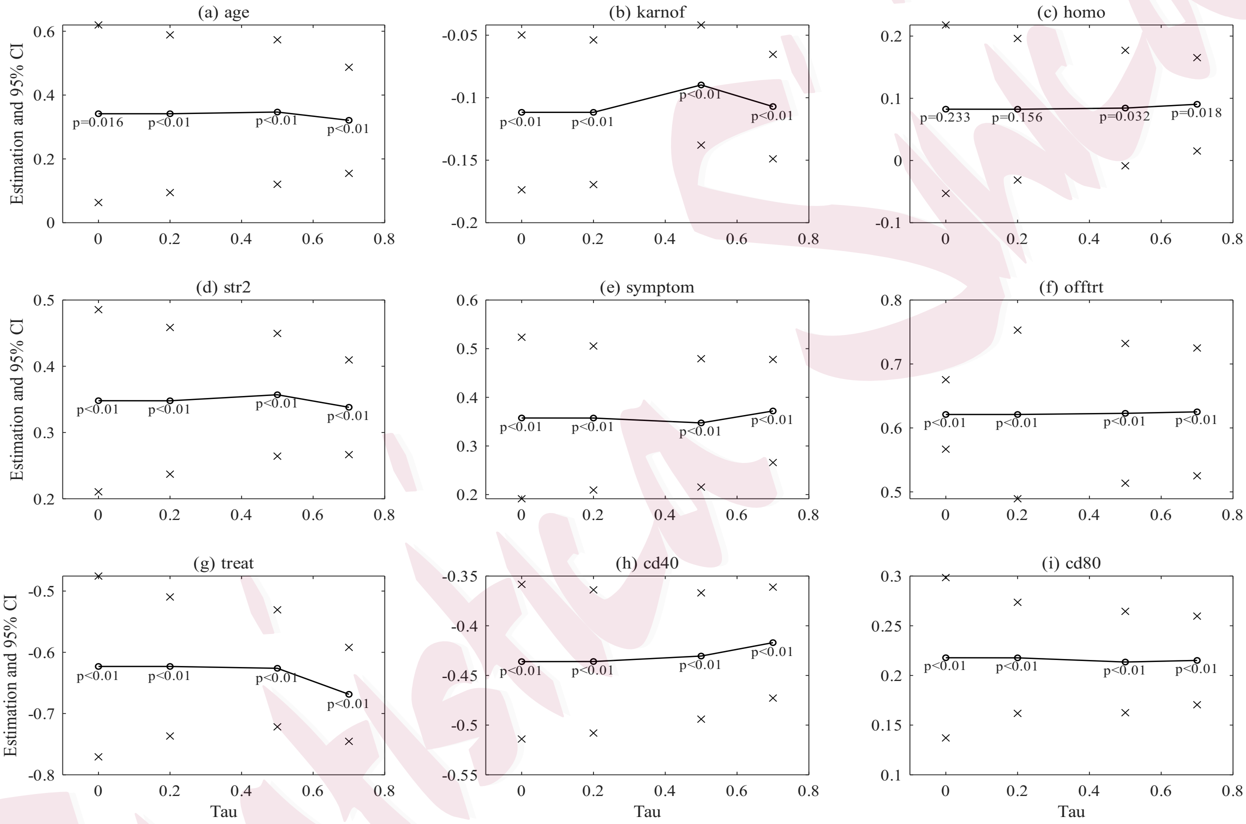

Figure 2: Plots of regression coefficients estimates (dots) and their corresponding 95% confidence intervals at τ values of 0, 0.2, 0.5, and 0.7.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

Figure 2 shows the regression coefficient versus τ values; each plot contains at least one significant coefficient. The plots also include the corresponding 95% confidence intervals (CIs), suggesting the following significant ($p < 0.05$) predictors of institutionalization regardless of the τ values: age, Karnofsky score, antiretroviral history, symptomatic indicator, treatment indicator, indicator of off-treatment before 96 ± 5 weeks, CD4 T cell count (cells/mm^3) at baseline, and CD8 T cell count (cells/mm^3) at baseline. Other predictors varied in statistical significance across Kendall's τ values. The effects of homosexual activity become marginally significant, which changes from nonsignificant ($p = 0.23$ and 0.16, when $\tau = 0$ and 0.2, respectively) to significant ($p = 0.03$ and 0.2 for $\tau = 0.5$ and 0.7, respectively).

6. Discussion

This article considered a computationally feasible variable selection method for PH models with dependent censoring. The associated likelihood function is nonconcave, involving unspecified baseline hazards that cannot be canceled due to the event and censoring times being correlated. To tackle these issues, we adopted the MIC method (Su et al., 2016) with copulas for its smooth formulation. The MIC method achieves sparse estimation by minimizing an approximated BIC. We adopted an efficient modified Newton-MI algorithm to estimate the baseline hazard and regression coefficients, which can be conveniently implemented in MATLAB. The consistency and asymptotic normality of the parameter estimators are established. Simulation results demonstrated that the proposed method performs satisfactorily. An application to the ACTG 175 dataset was provided to illustrate the utility of our method.

In the current research, we employed the hyperbolic tangent function (Su et al., 2016) to approximate the L_0 penalty for popular survival models. The MIC penalty has a simpler form and can provide a better and more efficient way along with nice properties than conventional ones. We also developed an iterative algorithm to effectively implement the proposed procedure. As there is no need to compute the inverse of the Hessian matrix, the proposed algorithm also works for large-scale problems. In addition, our method can be extended in several directions. First, our approach can be applied to other survival models, such as additive hazards models (Lin and Ying, 1994) and the accelerated failure time models. Second, variable selection for PH models with latent variables (Pan et al., 2015) can be considered in this framework. Third, the proposed method requires the variable dimension p to be less than the sample size n. The proposed method works well only in this situation because of its intricacy. Extending the current approach to deal with the high-dimensional case of $p \gg n$ is interesting.

Supplementary Material

The Supplementary Material includes the proofs of Theorems 1 to 3, Algorithm 1, an extended BIC for tuning parameter selection, and additional numerical results.

Acknowledgments

The authors are thankful to the editor, the associate editor, and two anonymous reviewers for their valuable comments and suggestions, which have substantially improved the article. This research is supported by the National Statistical Scientific Research Project of China (No. 2022LZ28), the Natural Science Foundation of China (No. 12271060), and the Research Grant Council of the Hong Kong Special Administration Region (No. 14303622).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE. T. Automat. Contr.* **19**, 716–723.

Brodaty, H., Woodward, M., Boundy, K., Ames, D. and Balshaw, R. (2011). Patients in australian memory clinics: baseline characteristics and predictors of decline at six months. *Int. Psychogeriatr.* **23**, 1086–1096.

Brodaty, H., Connors, M. H., Xu, J., Woodward, M., Ames, D. and PRIME Study Group. (2014). Predictors of institutionalization in dementia: a three year longitudinal study. *J. Alzheimers Dis.* **40**, 221–226.

Chen, Y. H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72**, 235–251.

Chen, X., Hu, T. and Sun, J. (2017). Sieve maximum likelihood estimation for the proportional hazards model under informative censoring. *Comput. Statist. Data Anal.* **112**, 224–234.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **34**, 187–220.

Dicker, L., Huang, B. and Lin, X. (2013). Variable selection and estimation with the seamless-L0 penalty. *Stat. Sinica* **23**, 929–962.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–499.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0227

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

Fan, J. and Li, R. (2002). Variable Selection for Cox's Proportional Hazards Model and Frailty Model. *Ann. Statist.* **30**, 74–99.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Stat.* **7**, 397–416.

Fu, W. J. (2005). Nonlinear GCV and quasi-GCV for shrinkage models. *J. Stat. Plan. Infer.* **131**, 333–347.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New Engl. J. Med.* **335**, 1081–1090.

Ha, I. D., Lee, M., Oh, S., Jeong, J. H., Sylvester, R. and Lee, Y. (2014). Variable selection in subdistribution hazard frailty models with competing risks data. *Stat. Med.* **33**, 4590–4604.

Han, D., Liu, L., Su, X., Johnson, B. and Sun, L. (2019). Variable selection for random effects two-part models. *Stat. Methods Med. Res.* **28**, 2697–2709.

Huang, X. and Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics* **58**, 510–520.

Huang, X. and Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics* **64**, 1090–1099.

Jo, J. H., Gao, Z., Jung, I., Song, S. Y., Ridder, G. and Moon, H. R. (2023). Copula graphic estimation of the survival function with dependent censoring and its application to analysis of pancreatic cancer clinical trial. *Stat. Methods Med. Res.* **32**, 944–962.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28**, 1356–1378.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.

Ma, J. (2010). Positively constrained multiplicative iterative algorithm for maximum penalized likelihood tomographic reconstruction. *IEEE T. Nucl. Sci.* **57**, 181–192.

Ma, J., Heritier, S. and Lô, S. N. (2014). On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. *Comput. Statist. Data Anal.* **74**, 142–156.

Pan, D., He, H., Song, X. Y. and Sun, L. Q. (2015). Regression analysis of additive hazards model with latent variables. *J. Amer. Statist. Assoc.* **110**, 1148–1159.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104**, 735–746.

Reeder, H. T., Lu, J. and Haneuse, S. (2023). Penalized estimation of frailty-based illness-death models for semi-competing risks. *Biometrics* **79**, 1657–1669.

Scharfstein, D. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika* **89**, 617–634.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464 .

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.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox Model. *Stat. Med.* **16**, 385–395.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**, 256–262.

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.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 49–67.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **3**, 691–703.

Zhao, H., Wu, Q. W., Li, G. and Sun, J. G. (2020). Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *J. Amer. Statist. Assoc.* **115**, 204–216.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

Zhang, Y., Wang, J. and Zhang, W. (2024). The adaptive lasso and its oracle properties. *Statist. Theor. Relat.* **8**, 211–231.

Zili Liu

School of Mathematics and Statistics, Central South University, Changsha, 410083, China

E-mail: (zili_liu@csu.edu.cn)

Hong Wang

School of Mathematics and Statistics, Central South University, Changsha, 410083, China

E-mail: (wh@csu.edu.cn)

Chunjie Wang

School of Mathematics and Statistics, Changchun University of Technology, Changchun, China

E-mail: (wangchunjie@ccut.edu.cn)

Xinyuan Song

Department of Statistics, Chinese University of Hong Kong, Hong Kong, China

E-mail: (xysong@sta.cuhk.edu.hk)