Published in final edited form as:
Lifetime Data Anal. 2010 April ; 16(2): 231–249. doi:10.1007/s10985-009-9139-z.

A copula model for bivariate hybrid censored survival data with application to the MACS study

Suhong Zhang,
Division of Biostatistics, Edwards Lifesciences, One Edwards Way, Irvine, CA 92612, USA suhong.zhang@edwards.com

Ying Zhang,
Department of Biostatistics, University of Iowa, C22 GH, 200 Hawkins Drive, Iowa City, IA 52242, USA ying-j-zhang@uiowa.edu

Kathryn Chaloner, and
Department of Biostatistics, University of Iowa, C22 GH, 200 Hawkins Drive, Iowa City, IA 52242, USA kathryn-chaloner@uiowa.edu

Jack T. Stapleton
Department of Internal Medicine, University of Iowa and Iowa City VA Medical Center, SW54-15 GH, 200 Hawkins Drive, Iowa City, IA 52242, USA jack-stapleton@uiowa.edu

Abstract

A copula model for bivariate survival data with hybrid censoring is proposed to study the association between survival time of individuals infected with HIV and persistence time of infection with an additional virus. Survival with HIV is right censored and the persistence time of the additional virus is subject to interval censoring case 1. A pseudo-likelihood method is developed to study the association between the two event times under such hybrid censoring. Asymptotic consistency and normality of the pseudo-likelihood estimator are established based on empirical process theory. Simulation studies indicate good performance of the estimator with moderate sample size. The method is applied to a motivating HIV study which investigates the effect of GB virus type C (GBV-C) co-infection on survival time of HIV infected individuals.

Keywords

Association measure; Bivariate survival model; Copula; Current status data; Kendall's τ; Right censored data; Empirical process

1 Introduction and motivating example

This paper was motivated by the investigation of the association between survival time among HIV-infected subjects and co-infection with an additional apparently harmless virus named GB Virus Type C (or GBV-C). Several recent studies suggest that persistent co-infection of GBV-C is associated with prolonged HIV survival (for example, Xiang et al. 2001; Tillmann et al. 2001; Williams et al. 2004; Zhang et al. 2006), while this beneficial association was not significant in other studies (Toyoda et al. 1998; Birk et al. 2002).

Correspondence to: Ying Zhang.

Among all these studies, the Multicenter AIDS Cohort Study (MACS, Williams et al. 2004) is the most comprehensive study to date. It began to recruit subjects at risk for HIV infection from 1984, a time close to the beginning of the AIDS epidemic. For each subject, blood samples were taken and stored every 6 months. When diagnostic testing for HIV subsequently became available, seroconverters were identified through retrospective testing of the stored samples. Later, for a selected subset of seroconverters, two samples of stored blood were tested for GBV-C infection: one sample at 12–18 months after the subject's first positive HIV test (HIV onset), and the second was a sample at 4.5–6 years after seroconversion. The analysis conducted in Williams et al. (2004) treated all HIV survival times as right censored at January 1, 1996 to avoid confounding with the use of highly active HIV therapy that became available in 1996. They found that persistent GBV-C infection was significantly associated with prolonged survival among HIV-positive subjects at the late time (4.5–6 years after HIV onset), but not at the early time (12–18 months after HIV onset).

All previous studies compared the Kaplan-Meier survival curves between HIV-infected subjects with and without GBV-C infection at a specified time using the log-rank test. However, GBV-C viremia may clear over time and GBV-C persistence time varies among subjects. As a consequence, if GBV-C persistence time plays an essential role in its association with HIV survival then time to GBV-C clearance needs to be included in any comparison. This motivated the need to model GBV-C persistence time, rather than the status at a single time.

The use of Cox regression model with GBV-C status treated as a time-dependent covariate is not possible in this MACS data set. The Cox model requires that GBV-C status be known throughout the time during the study (Kalbfleisch and Prentice 2002, p. 200), but GBV-C status in the MACS study is only known at baseline and one another follow-up time.

In this paper we propose a bivariate survival model to adjust for the GBV-C persistence time since co-infection (time from HIV onset to the clearance of GBV-C). The GBV-C diagnostic test at the time close to HIV seroconversion is treated as the baseline GBV-C status, and the test at the second observation time provides current status data on GBV-C persistence time. Current status data, or interval censoring case 1 data, is a special case of interval censoring when it is only feasible to know whether or not an event (clearance) has occurred at a monitoring time (Groeneboom and Wellner 1992).

Bivariate and multivariate survival data have been studied extensively in the statistical literature. Liang et al. (1995) and Oakes (2000) reviewed some recent developments for analysis of multivariate failure time data. Copula based survival models are considered, for example, by Hougaard (1989), Oakes (1989), Shih and Louis (1995) and Wang and Ding (2000). Shih and Louis (1995) examined the association of the bivariate data that are both subject to right censoring, through a two-stage semiparametric estimation procedure. At the first stage in their procedure, the marginal survival functions are estimated consistently by nonparametric maximum likelihood estimators. At the second stage, a dependency structure is imposed by using a copula model, and the nonparametric maximum likelihood estimators of the two marginal survival functions are substituted into the likelihood function to form a pseudo-likelihood, then the association parameter is estimated through a pseudo-likelihood approach. Wang and Ding (2000) proposed a parallel two-stage semiparametric method for the bivariate current status data. In both papers, they showed that the proposed estimators of the association measure converge in distribution to normal random variables with the $n^{1/2}$ rate without demonstrating the consistency first which is, however, required in the proof of asymptotic normality.

In this paper, we model the association of bivariate event times using a copula model and estimate the association parameter through the two-stage procedure as well. We focus specifically on the data structure where one of the paired event time data is right-censored and the other is observed as current status data, as observed in the MACS study. Our main goal in this paper is to develop an inference procedure to study the association of bivariate survival data with this type of censoring structure and to apply the proposed method to investigate the association between HIV survival and GBV-C persistence time.

The rest of this paper is organized as follows. Section 2 introduces a theoretical model and describes a two-stage semiparametric estimation procedure for the association parameter. Section 3 states asymptotic properties of the association parameter estimator. Section 4 presents simulation studies. Section 5 applies the proposed estimation method to the MACS GBV-C study. Finally, Section 6 summarizes the method with some remarks. Technical details are provided in the Appendix.

2 Likelihood and estimation method

In what follows the usual cumulative distribution function is defined as $F(t) = P(T \leq t)$ and the corresponding survival function is defined as $S(t) = P(T > t) = 1 - F(t)$.

Let T_1^0 be the HIV survival time and T_2^0 be the GBV-C persistence time. Assume the distributions of T_1^0 and T_2^0 are continuous. Let S_j and F_j, $j = 1, 2$ be the survival function and distribution function of T_j^0, respectively. Denote $F(t_1, t_2)$ and $S(t_1, t_2)$ the joint distribution function and survival function of $\left(T_1^0, T_2^0\right)$, respectively. We propose to model the joint survival function $S(t_1, t_2)$ by the one-parameter Archimedean copula C_α:

$$C_\alpha:[0,1]^2 \to [0,1] \quad \text{that satisfies} \quad S_\alpha(t_1,t_2) = C_\alpha(S_1(t_1), S_2(t_2)).$$

The joint distribution function $Fa(t_1, t_2)$ can therefore be expressed as $F_\alpha(t_1, t_2) = 1 - S_1(t_1) - S_2(t_2) + S_\alpha(t_1, t_2)$.

Examples of various one-parameter Archimedean copula models are discussed in Nelsen (2006). As Kendall's τ is related to the Copula by $\tau = 4E\{C_\alpha(u, v)\} - 1$ (Nelsen 2006), the parameter α is naturally linked to the association between the two random variables with the marginal survival functions given by S_1 and S_2, respectively. Therefore, the inference for the association between the two event times can be made through the inference about α.

We consider bivariate survival data with hybrid censoring in which T_1^0 is right censored by a random variable C_1 and T_2^0 is subject to interval censoring case 1 by a random monitoring time C_2. Suppose we have collected a random sample of $(T_{1i}, T_{2i}, \Delta_{1i}, \Delta_{2i})$, $i = 1, 2, \ldots, n$, from a distribution with density function $f(t_1, t_2, \delta_1, \delta_2)$, where $T_{1i} = T_{1i}^0 \wedge C_{1i}$ and $T_{2i} = C_{2i}; \Delta_{1i} = 1_{[T_{1i}^0 \leq C_{1i}]}$ and $\Delta_{2i} = 1_{[T_{2i}^0 \leq C_{2i}]}$. We consider the scenario of independent and non-informative censoring, i.e., $\left(T_1^0, T_2^0\right)$ are jointly independent of (C_1, C_2), and the distribution of (C_1, C_2) is non-informative to any parameters in the joint distribution of $\left(T_1^0, T_2^0\right)$. We also denote $G_i(t)$ the marginal distribution function of C_i with density function $g_i(t)$, for $i = 1, 2$.

The density function $f(t_1, t_2, \delta_1, \delta_2)$ can be explicitly written for four distinct cases with respect to Lebesgue measure. Combining the four cases and discarding the parts that are

non-informative to the joint distribution of T_1^0 and T_2^0, we can derive the likelihood for n independently and identically distributed observations.

Let $C_{1\alpha}(u,v) = \frac{\partial}{\partial u} C_\alpha(u,v)$. Given the marginal survival functions S_1 and S_2, the likelihood for α, omitting parts that are irrelevant in estimating α, is

$$\begin{aligned} L(\alpha, S_1, S_2; data) &= \prod_{i=1}^{n} [1 - C_{1\alpha}(S_1(t_{1i}), S_2(t_{2i}))]^{\delta_{1i}\delta_{2i}} [C_{1\alpha}(S_1(t_{1i}), S_2(t_{2i}))]^{\delta_{1i}(1-\delta_{2i})} \\ &\times [S_1(t_{1i}) - C_\alpha(S_1(t_{1i}), S_2(t_{2i}))]^{(1-\delta_{1i})\delta_{2i}} [C_\alpha(S_1(t_{1i}), S_2(t_{2i}))]^{(1-\delta_{1i})(1-\delta_{2i})}. \end{aligned} \quad (1)$$

A two-stage maximum pseudo-likelihood estimation approach is developed to estimate α. The first stage involves the estimation of marginal survival functions for censored data. The marginal survival function S_1 is estimated by the Kaplan-Meier estimator $\hat{S}_1$ and S_2 is estimated by the nonparametric maximum likelihood estimator $\hat{S}_2$, using the Convex Minorant Algorithm described by Groeneboom and Wellner (1992).

At the second stage, $\hat{S}_1(t)$ and $\hat{S}_2(t)$ are substituted into the likelihood (1), the resulting pseudo-likelihood is then maximized with respect to α. The maximum pseudo-likelihood estimator $\widehat{\alpha}_n$ is the solution to the pseudo score equation:

$$U_\alpha\left(\alpha, \widehat{S}_1, \widehat{S}_2; data\right) = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} l\left(\alpha, \widehat{S}_1(t_{1i}), \widehat{S}_2(t_{2i}), \delta_{1i}, \delta_{2i}\right) = 0, \quad (2)$$

where

$$\begin{aligned} l\left(\alpha, \widehat{S}_1(t_1), \widehat{S}_2(t_2), \delta_1, \delta_2\right) &= \delta_1\delta_2 \log\left(1 - C_{1\alpha}\left(\widehat{S}_1(t_1), \widehat{S}_2(t_2)\right)\right) \\ &+ \delta_1(1-\delta_2) \log C_{1\alpha}\left(\widehat{S}_1(t_1), \widehat{S}_2(t_2)\right) \\ &+ (1-\delta_1)\delta_2 \log\left(\widehat{S}_1(t_1) - C_\alpha\left(\widehat{S}_1(t_1), \widehat{S}_2(t_2)\right)\right) \\ &+ (1-\delta_1)(1-\delta_2) \log C_\alpha\left(\widehat{S}_1(t_1), \widehat{S}_2(t_2)\right). \end{aligned} \quad (3)$$

Note that the pseudo likelihood approach was previously adopted by Shih and Louis (1995) in an association study of bivariate right censored data and by Wang and Ding (2000) in a study of association between two event times with both subject to interval censoring case 1 (Groeneboom and Wellner 1992).

3 Asymptotic properties of the maximum pseudo-likelihood estimator $\widehat{\alpha}_n$

Let T_1 and T_2 take values on $[0, t_{01}] \times [0, t_{02}]$, where $t_{01} = \sup\{t: P(T_1 > t, C_1 > t) > 0\}$ and $t_{02} = \sup\{t: P(C_2 > t) > 0\}$. Suppose α is in an open set A in the real line. Denote D a universal constant throughout the rest of technical development.

For the brevity of presentation, we define the following notations:

$$
\begin{aligned}
V_{\alpha}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) &= \frac{\partial}{\partial \alpha} l(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) \\
V_{\alpha^2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) &= \frac{\partial^2}{\partial \alpha^2} l(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) \\
V_{\alpha,1}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) &= \frac{\partial^2}{\partial \alpha \partial u} l(\alpha, u, S_2(t_2), \delta_1, \delta_2)|_{u=S_1(t_1)} \\
V_{\alpha,2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) &= \frac{\partial^2}{\partial \alpha \partial v} l(\alpha, S_1(t_1), v, \delta_1, \delta_2)|_{v=S_2(t_2)} \\
V_{\alpha^2,1}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) &= \frac{\partial^3}{\partial \alpha^2 \partial u} l(\alpha, u, S_2(t_2), \delta_1, \delta_2)|_{u=S_1(t_1)} \\
V_{\alpha^2,2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) &= \frac{\partial^3}{\partial \alpha^2 \partial v} l(\alpha, S_1(t_1), v, \delta_1, \delta_2)|_{v=S_2(t_2)} \\
V_{\alpha,1^2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) &= \frac{\partial^3}{\partial \alpha \partial u^2} l(\alpha, u, S_2(t_2), \delta_1, \delta_2)|_{u=S_1(t_1)} \\
V_{\alpha,1,2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) &= \frac{\partial^3}{\partial \alpha \partial u \partial v} l(\alpha, u, v, \delta_1, \delta_2)|_{u=S_1(t_1), v=S_2(t_2)} \\
V_{\alpha,2^2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) &= \frac{\partial^3}{\partial \alpha \partial v^2} l(\alpha, S_1(t_1), v, \delta_1, \delta_2)|_{v=S_2(t_2)}
\end{aligned}
$$

To study the asymptotic properties of $\widehat{\alpha}_n$, we need the following regularity conditions. Some of the conditions are related to the smoothness of the copula models and the likelihood.

A1 $l(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2)$ is three-time differentiable with respect to α on $[0, t_{01}] \times [0, t_{02}]$, for each $\alpha \in A$, and all derivatives are continuous and uniformly bounded by some constant D.

A2 $V_{\alpha,1}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2)$, $V_{\alpha,2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2)$, $V_{\alpha^2,1}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2)$, $V_{\alpha^2,2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2)$, $V_{\alpha,1^2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2)$, $V_{\alpha,1,2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2)$, and $V_{\alpha,2^2}(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2)$ exist and are uniformly bounded by some constant D on $[0, t_{01}] \times [0, t_{02}]$, for all $\alpha \in A$ and survival functions S_1 and S_2.

A3 For each $\alpha \in A$, $0 < \mathrm{E}_{\alpha}[V_{\alpha}(\alpha, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2)]^2 < \infty$.

A4 F_2 and G_2 are absolutely continuous with respect to each other.

A5 $(\psi_2/g_2) \circ S_2^{-1}$ is bounded and Lipschitz on [0, 1], where ψ_2 is the derivative of the influence curve $IC_2(t_2)$, defined by

$$IC_2(t_2) = -\int_0^{t_2}\int_0^{t_{01}} V_{\alpha,2}(\alpha_0, S_1(\tau_1), S_2(\tau_2), \delta_1, \delta_2)\, dP(\tau_1, \tau_2, \delta_1, \delta_2).$$

A6 S_2, g_2 and ψ_2 satisfy

$$\int_0^{t_{02}} \frac{S_2(t_2)(1 - S_2(t_2))}{g_2(t_2)} \psi_2(t_2)\, dt_2 < \infty.$$

Remarks

Conditions (A1) and (A2) require the log likelihood to be differentiable with respect to the unknown parameters. These conditions can be easily but tediously verified for Archimedean copulas. Condition (A3) indicates the log likelihood has finite nonzero information about α when the marginal survival functions are known which is usually required in parametric maximum likelihood theory. Conditions (A4)–(A6) are the regularity conditions given by Huang and Wellner (1995) in studying the asymptotic normality of linear functionals of the nonparametric maximum likelihood estimator of S_2 with current status data. These regularity conditions are generally mild for applications.

The following two lemmas are important to study asymptotic properties of $\widehat{\alpha}_n$.

Lemma 1 *Let* $\mathcal{F}_j = \{f : f \text{ is a survival function on } [0, t_{0j}]\}$, $j = 1, 2$, *and the class* $\mathcal{G}_{\mathcal{F}} = \{V_{\alpha,1}(\alpha, f_1(t_1), f_2(t_2), \delta_1, \delta_2); f_j \in \mathcal{F}_j, j=1,2\}$. *Let P denote the probability measure of* (T_1, T_2, Δ_1, Δ_2), *then under condition (A1)–(A2),* $\mathcal{G}_{\mathcal{F}}$ *is a P-Glivenko-Cantelli class, for all* $\alpha \in A$.

Lemma 2 *Let* $\mathcal{F}_j = \{f : f \text{ is a survival function on } [0, t_{0j}]\}$, $j = 1, 2$ *and the class* $\mathcal{H}_{\mathcal{F}} = \{V_\alpha(\alpha, f_1(t_1), f_2(t_2), \delta_1, \delta_2) - V_\alpha(\alpha, S_1(t_1), S_2(t_2), \delta_1, \delta_2) : f_j \in \mathcal{F}_j, j=1,2\}$. *Let P denote the probability measure of* (T_1, T_2, Δ_1, Δ_2), *then under condition (A1)–(A2),* $\mathcal{H}_{\mathcal{F}}$ *is a P-Donsker Class, for all* $\alpha \in A$.

Based on these two lemmas, the maximum pseudo-likelihood estimator $\widehat{\alpha}_n$ can be shown consistent and asymptotically normally distributed. The results are summarized in the following two theorems.

Theorem 1 *Assume that the joint distribution of* (T_1^0, T_2^0) follows an Archimedean copula model with the true association parameter $\alpha = \alpha_0$. *Under the regularity conditions (A1)–(A2),* $\widehat{\alpha}_n \xrightarrow{p} \alpha_0$ *as* $n \to \infty$.

Theorem 2 *Under the regularity conditions (A1)–(A6),* $\sqrt{n}(\widehat{\alpha}_n - \alpha_0) \xrightarrow{d} N(0, \sigma^2)$, *where*

$$\sigma^2 = \frac{Var(Q(T_1, T_2, \Delta_1, \Delta_2; \alpha_0, S_1, S_2))}{W^2(\alpha_0, S_1, S_2)}$$

with

$$\begin{aligned} W(\alpha_0, S_1, S_2) &= -\int [V_\alpha(\alpha_0, S_1(t_1), S_2(t_2), \delta_1, \delta_2)]^2 dP(t_1, t_2, \delta_1, \delta_2) \\ Q(T_1, T_2, \Delta_1, \Delta_2; \alpha_0, S_1, S_2) &= V_\alpha(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2) + I_1(T_1, \Delta_1; \alpha_0) - \tilde{l}(T_2, \Delta_2; S_2, G_2, \psi_2), \end{aligned}$$

in which

$$\begin{aligned} &I_1(T_1, \Delta_1; \alpha_0) \\ &= \int_0^{t_{01}} \int_0^{t_{02}} M_{\alpha,1}(\alpha_0, S_1(\tau_1), S_2(\tau_2)) f(\tau_1, \tau_2) I_1^0(T_1, \Delta_1)(\tau_1) d\tau_1 d\tau_2 \quad \text{and} \quad \tilde{l}(T_2, \Delta_2; S_2, G_2, \psi_2) \\ &= -[\Delta_2 - (1 - S_2(T_2))] \frac{\psi_2(T_2)}{g_2(T_2)} I[g_2(T_2) > 0], \end{aligned}$$

where

$$M_{\alpha,1}(\alpha_0, S_1(t_1), S_2(t_2)) = -E\{V_{\alpha,1}(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2) | T_1 = t_1, T_2 = t_2\}$$

and

$$I_1^0(T_1, \Delta_1)(t_1) = -S_1(t_1) \left\{ \int_0^{t_1} \frac{1}{P(T_1 \geq u)} dN_1(u) - \int_0^{t_1} \frac{I[T_1 \geq u]}{P(T_1 \geq u)} d\Lambda_1(u) \right\}.$$

Here $N_1(u)$ *is defined as* $I[T_1 \le u, \Delta_1 = 1]$ *and* Λ_1 *is the cumulative hazard function of* T_1^0.

The proofs of these lemmas and theorems are provided in the Appendix.

4 Simulation studies

Simulation studies are conducted to evaluate the finite sample performance of the proposed method. A Gumbel copula, a special case of Archimedean copulas, defined by

$$C_\alpha(u,v) = exp\left\{-\left[(-\log u)^\alpha + (-\log v)^\alpha\right]^{1/\alpha}\right\}, \quad \alpha \ge 1, \quad 0 \le u, v \le 1$$

is used to generate the bivariate event time data in which the two marginal distributions are both assumed to be exponential with unit rate 1. For the Gumbel copula, a larger α corresponds to a stronger positive association and $\alpha = 1$ corresponds to the case that the two event times are independent.

A sample of bivariate copula random variables is generated based on their conditional distribution function. Suppose that the joint distribution of the bivariate data $\left(T_1^0, T_2^0\right)$ is $C_\alpha(1 - \exp(-t_1), 1 - \exp(-t_2))$. We generate $\left(T_1^0, T_2^0\right)$ through the following steps:

– Generate two independent uniform (0, 1) random variables u, w.

– Set $w = P(V \le v | U = u) = \partial C_\alpha(u, v)/\partial u$, solve for v.

– Set $T_1^0 = -\log(1-u)$, $T_2^0 = -\log(1-v)$.

Meanwhile, a sample of bivariate censoring times (C_1 and C_2) are each independently drawn from a uniform distribution on [0, 2.3]. In this setting, about 50% of T_1^0 is right censored by C_1, and about 50% of T_2^0 is subject to interval censoring case 1 by C_2 as well.

Kendall's τ is chosen as a global association measure. For the Gumbel copula, $\tau = 1 - 1/\alpha$. Three different values of α are set such that the corresponding Kendall's τ is 0.25, 0.5, and 0.75. For each value of α, we conduct Monte-Carlo simulations with 1,000 replications for sample size $n = 50$, 100, 200 and 400, respectively.

For each of the 1,000 simulations, Wald confidence interval is constructed based on the asymptotic normality, in which the standard error of $\widehat{\alpha}_n$ is computed using 200 bootstrap resamples. The empirical estimate of the coverage probability is obtained based on the Wald confidence interval over 1,000 replications.

Table 1 summarizes the simulation results for the two-stage pseudo-likelihood estimator. It provides results for estimation bias, Monte-Carlo standard deviation of 1,000 replicates as the empirical standard error (ese), mean of bootstrap standard error (bse), and 95% empirical coverage probability (ecp).

As sample size increases, for a wide range of α, the biases of both $\widehat{\alpha}_n$ and $\widehat{\tau}_n$ decreases considerably, so do the Monte-Carlo standard deviation and bootstrap standard error. In addition, when sample size increases, the empirical coverage probability converges to the nominal level and the Monte-Carlo standard deviation and the mean of bootstrap standard error tend to get closer.

With same sample size, the stronger the dependency, the bigger the bias and the standard error for the estimator $\widehat{\alpha}_n$, as greater variations are usually expected for larger values.

Therefore, to preserve high efficiency, a larger sample size is desired to achieve reasonable performance of $\widehat{\alpha}_n$ when a strong association exists. Interestingly, we observe that the standard deviation of $\widehat{\tau}_n$ decreases as the association becomes stronger. This may be explained by the standard delta method which implies that $\sigma_{\widehat{\tau}_n} \approx \frac{\sigma_{\widehat{\alpha}_n}}{\alpha^2}$ when sample size is large. The simulations demonstrate that this relationship approximately holds when $n \geq 200$. We also note that the average of bootstrap standard error of estimated Kendall's τ is very close to the Monte-Carlo standard deviation when $n \geq 100$, particularly if the association is not strong. This may imply the inference about the Kendall's τ will be reasonably good when $n \geq 100$.

In addition to compute the proposed two-sage maximum pseudo-likelihood estimator $\widehat{\alpha}_n$, we also compute $\tilde{\alpha}_n$, the maximum likelihood estimator when the two marginal distributions are completely specified. The latter estimator serves as a benchmark to evaluate the performance of the maximum pseudo-likelihood estimator. Table 2 gives the results of $\tilde{\alpha}_n$ and $\tilde{\tau}_n$, the maximum likelihood estimators of α and τ, respectively, when the two marginal survival functions are known. The maximum likelihood estimators perform better than the proposed maximum pseudo-likelihood estimators, as expected, but their differences are substantially reduced when sample size increases, say $n \geq 200$. The small difference between the two estimators assures us the use of two-stage pseudo-likelihood estimation procedure, for which we gain the advantage of having flexibility by not modeling the marginal distributions but maintain high estimation efficiency with reasonable sample size.

5 Application to the motivating example

We apply the proposed method to the sub-cohort of MACS from Williams et al. (2004) to study the association of GBV-C persistence time and HIV survival. MACS consists of gay men who were enrolled between 1984 and 1990 and whose blood samples were obtained every 6 months and tested retrospectively when a test for HIV became available. The sub-cohort includes 271 subjects from MACS who were initially HIV negative when they entered the study but HIV positive during the follow ups. Since the visits were scheduled every 6 months, the seroconversion time is known to be within a six-month window. Seroconversion time is imputed as the midpoint between the last seronegative visit and the first seropositive visit. All 271 subjects were evaluated at 12–18 months after HIV seroconversion for the evidence of GBV-C infection and a subgroup of 138 patients were re-examined 4.5–6 years after HIV seroconversion. The study only included data collected before Jan 1, 1996 to avoid the impact of the use of highly active antiretroviral therapy.

Williams et al. (2004) compared the Kaplan-Meier curves for the survival time of the HIV subjects with and without GBV-C co-infection at disease onset and found no significant difference at level 0.05. Here we consider the association between GBV-C persistence time and HIV survival among people who were co-infected with both HIV and GBV-C at HIV onset. HIV survival is defined as the time from seroconversion to death, and GBV-C persistence time is defined as the time from HIV seroconversion to GBV-C clearance for the subjects with GBV-C positive at HIV onset. Previous clinical studies and lab studies suggest that the re-infection of GBV-C is very rare among people who have already infected with HIV. So we assume that the HIV subjects who were co-infected with GBV-C would not be re-infected once they lose it.

In our analysis, we treat the GBV-C status evaluated at 12–18 months as the baseline GBV-C information to select a subsample of HIV patients who are assumed to be co-infected with GBV-C at HIV onset. The GBV-C status evaluated at the second time after HIV seroconversion presents the current status data for GBV-C persistence time. The Gumbel copula is used for the bivariate distribution of HIV survival and GBV-C persistence times.

The bootstrap standard error based on 1,000 resamples with replacement was used to estimate the standard error of the estimated association parameter and to construct the Wald confidence interval. There are 61 subjects who were GBV-C positive at the first visit, and GBV-C status at the late visit were known and evaluated before January 1, 1996. In order to use as many data as possible, we define the current status of GBV-C co-infection for the subjects whose late observations on GBV-C were unavailable before January 1, 1996 as follows: (i) for those whose second GBV-C test were negative and evaluated after January 1, 1996, their GBV-C persistence times were right censored at the first visit ($n = 2$); (ii) for those whose second GBV-C test were positive and evaluated after January 1, 1996, their GBV-C persistence times were right censored at January 1, 1996 ($n = 7$); and (iii) for those whose second GBV-C test were missing, their GBV-C persistence times were right censored at the first visit ($n = 37$). Therefore, we have a total of 107 subjects for analysis. Table 3 presents the results when all the subjects who were GBV-C positive at the first visit are included. The maximum pseudo-likelihood estimate of Kendall's τ is $\widehat{\tau}_n = 0.3685$ with an 95% confidence interval being [0.1988, 0.5383] using the asymptotic normality or [0.2114, 0.5533] using the bootstrap method. The result indicates that GBV-C persistence time is moderately associated with increased survival among HIV and GBV-C co-infected individuals.

6 Final remarks

This manuscript proposes a method for assessing the association between two random variables which are subject to different censoring schemes: one is right censored and the other is observed as current status data. The asymptotic properties of the estimator of association parameter, including consistency and asymptotic normality, are established under mild technical assumptions. Although the asymptotic variance of the estimator has a complicated form and is difficult to estimate directly, the ordinary bootstrap method provides a practical and efficient way to estimate the standard error.

Our simulation results suggest that the proposed method works well for moderate sample size and has the advantage of allowing for flexibility in the marginal distributions. Moreover, our numerical study shows that the proposed method is quite efficient compared to the full maximum likelihood approach in which the marginal distributions are given. It suggests that the efficiency loss from the pseudo-likelihood approach is not substantial.

Some copula functions, such as the Gumbel copula, are equivalent to the independent copula only when the association parameter takes its value on the boundary of the parameter space. It may result in failure of some regularity conditions and hence the likelihood theory cannot be easily developed which makes the test of the independence of bivariate event times problematic using the copula models. Several nonparametric tests of dependence have been developed for the bivariate censored data (Oakes 1982; Shih and Louis 1996; Hsu and Prentice 1996; Ding and Wang 2004). A nonparametric test procedure to test the dependence between HIV survival and GBV-C persistence time needs to be developed under the hybrid censoring considered in this paper.

A new study testing additional stored samples from MACS cohort is being planned. There will be considerably more power and precision using these additional time points in the analysis. With the new study of more GBV-C screening, GBV-C persistence time is interval censored, the method presented here will be extended to model the bivariate event data with one margin being subject to right censoring and the other being subject to interval censoring case 2. Other applications with time-varying covariates subject to interval censoring case 2 are readily available.

Acknowledgments

The authors wish to thank the Multicenter AIDS Cohort Study (MACS) for providing data. The MACS has centers located at: The Johns Hopkins Bloomberg School of Public Health (Joseph Margolick); Howard Brown Health Center and Northwestern University Medical School (John Phair); University of California, Los Angeles (Roger Detels); University of Pittsburgh (Charles Rinaldo); and Data Analysis Center (Lisa Jacobson). The authors are also thankful to the editors and two anonymous referees. Their insightful comments and suggestions greatly help improve this manuscript from an early version.

Appendix

This section provides proofs for the lemmas and theorems stated in Sect. 3. We use modern empirical process theory to justify our proofs. We denote $\int fdP$ by Pf and $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ by $\mathbb{P}_n f$.

Proof of Lemma 1 Since $\mathscr{F}_j$ consists of uniformly bounded monotone functions on the real line, by the Theorem 2.7.5 of **?**, for any $\varepsilon > 0$, for $j = 1, 2$, there exists a set of brackets:

$$\left[f_{j1}^L, f_{j1}^U\right], \left[f_{j2}^L, f_{j2}^U\right], \ldots, \left[f_{jN_j}^L, f_{jN_j}^U\right],$$

with $N_j \leq \exp(D/\varepsilon)$ and $\left(\int |f_{ji}^U - f_{ji}^L|^r dP\right)^{1/r} \leq \epsilon$ for any $1 \leq i \leq N_j$ and $r > 0$, such that for any $f_j \in \mathscr{F}_j$ and any $t_j \in \left[0, t_{0j}\right]$, $f_{jq_j}^L\left(t_j\right) \leq f_j\left(t_j\right) \leq f_{jq_j}^U\left(t_j\right)$ for some $1 \leq q_j \leq N_j$.

By (A2), $V_{\alpha,1}(\alpha, f_1(t_1), f_2(t_2), \delta_1, \delta_2)$ is continuous. We can then construct a set of brackets as follows: for any $i = 1, 2, \ldots, N_1$, $s = 1, 2, \ldots, N_2$ and for any $t_j \in [0, t_{0j}]$, we can find the unique maximum and minimum of $V_{\alpha,1}(\alpha, f_1(t_1), f_2(t_2), \delta_1, \delta_2)$ on the product set $\left[f_{1i}^L, f_{1i}^U\right] \times \left[f_{2s}^L, f_{2s}^U\right]$. Let

$$\begin{aligned}
\left(f_1^{L,(i,s)}(t_1), f_2^{L,(i,s)}(t_2)\right) &= \underset{\substack{f_1 \in \left[f_{1i}^L, f_{1i}^U\right] \\ f_2 \in \left[f_{2s}^L, f_{2s}^U\right]}}{\operatorname{argmin}} V_{\alpha,1}(\alpha, f_1(t_1), f_2(t_2), \delta_1, \delta_2) \\
\left(f_2^{U,(i,s)}(t_1), f_2^{U,(i,s)}(t_2)\right) &= \underset{\substack{f_1 \in \left[f_{1i}^L, f_{1i}^U\right] \\ f_2 \in \left[f_{2s}^L, f_{2s}^U\right]}}{\operatorname{argmax}} V_{\alpha,1}(\alpha, f_1(t_1), f_2(t_2), \delta_1, \delta_2)
\end{aligned}$$

and let

$$\begin{aligned}
V_{\alpha,1}^{L,(i,s)}(t_1, t_2, \delta_1, \delta_2) &= V_{\alpha,1}\left(\alpha, f_1^{L,(i,s)}(t_1), f_2^{L,(i,s)}(t_2), \delta_1, \delta_2\right) \\
V_{\alpha,1}^{U,(i,s)}(t_1, t_2, \delta_1, \delta_2) &= V_{\alpha,1}\left(\alpha, f_1^{U,(i,s)}(t_1), f_2^{U,(i,s)}(t_2), \delta_1, \delta_2\right).
\end{aligned}$$

The class $\mathscr{G}_{\mathscr{F}}$ is then covered by a set of $N_1 \times N_2$ brackets:

$$\left\{\left[V_{\alpha,1}^{L,(i,s)}(t_1, t_2, \delta_1, \delta_2), V_{\alpha,1}^{U,(i,s)}(t_1, t_2, \delta_1, \delta_2)\right] : i = 1, 2, \ldots, N_1, s = 1, 2, \ldots, N_2\right\}.$$

By (A2), $V_{\alpha,1}{}^2(\alpha, u, v, \delta_1, \delta_2)$ and $V_{\alpha,1,2}(\alpha, u, v, \delta_1, \delta_2)$ are bounded by some constant D, then $V_{\alpha,1}(\alpha, u, v, \delta_1, \delta_2)$ satisfies the Lipschitz condition with respect to u and v. It follows that:

$$\begin{aligned}\int|V_{\alpha,1}^{U,(i,s)}(t_1,t_2,\delta_1,\delta_2)-V_{\alpha,1}^{L,(i,s)}(t_1,t_2,\delta_1,\delta_2)|dP &= \int|V_{\alpha,1}\left(\alpha,f_1^{U,(i,s)}(t_1),f_2^{U,(i,s)}(t_2),\delta_1,\delta_2\right)-V_{\alpha,1}\left(\alpha,f_1^{L,(i,s)}(t_1),f_2^{L,(i,s)}(t_2),\delta_1,\delta_2\right)|dP\\ &\le \int\left[D|f_1^{U,(i,s)}(t_1)-f_1^{L,(i,s)}(t_1)|+D|f_2^{U,(i,s)}(t_2)-f_2^{L,(i,s)}(t_2)|\right]dP\\ &\le D\epsilon.\end{aligned}$$

This indicates that the preceding $N_1 \times N_2$ brackets are D_ε–brackets for $\mathscr{G}_\mathscr{F}$. It follows that, for any $\varepsilon > 0$, the bracketing number of class $\mathscr{G}_\mathscr{F}$ associated with $L_1(P)$ norm is bounded. By Theorem 2.4.1 of **?**, $\mathscr{G}_\mathscr{F}$ is a P-Glivenko-Cantelli class.

Proof of Lemma 2 Based on the similar technique used in the proof of Lemma 1, we can construct a set of $N_1 \times N_2$ brackets:

$$\left\{\left[V_\alpha^{L,(i,s)}(t_1,t_2,\delta_1,\delta_2)-V_\alpha(\alpha,S_1(t_1),S_2(t_2),\delta_1,\delta_2),V_\alpha^{U,(i,s)}(t_1,t_2,\delta_1,\delta_2)-V_\alpha(\alpha,S_1(t_1),S_2(t_2),\delta_1,\delta_2)\right]:i=1,2,\ldots,N_1,s=1,2,\ldots,N_2\right\}$$

which covers $\mathscr{H}_\mathscr{F}$.

By (A2), $V_{\alpha,1}(\alpha, u, v, \delta_1, \delta_2)$ and $V_{\alpha,2}(\alpha, u, v, \delta_1, \delta_2)$ are bounded by some constant D, then $V_\alpha(\alpha, u, v, \delta_1, \delta_2)$ satisfies the Lipschitz condition with respect to u and v. Also note that $(x + y)^2 = x^2 + y^2 + 2xy \le 2x^2 + 2y^2$, it follows that

$$\begin{aligned}\int\left(V_\alpha^{U,(i,s)}(t_1,t_2,\delta_1,\delta_2)-V_\alpha^{L,(i,s)}(t_1,t_2,\delta_1,\delta_2)\right)^2dP &= \int|V_\alpha\left(\alpha,f_1^{U,(i,s)}(t_1),f_2^{U,(i,s)}(t_2),\delta_1,\delta_2\right)-V_\alpha\left(\alpha,f_1^{L,(i,s)}(t_1),f_2^{L,(i,s)}(t_2),\delta_1,\delta_2\right)|^2dP\\ &\le \int\left[D|f_1^{U,(i,s)}(t_1)-f_1^{L,(i,s)}(t_1)|+D|f_2^{U,(i,s)}(t_2)-f_2^{L,(i,s)}(t_2)|\right]^2dP\\ &\le 2D^2\int|f_1^{U,(i,s)}(t_1)-f_1^{L,(i,s)}(t_1)|^2dP+2D^2\int|f_2^{U,(i,s)}(t_2)-f_2^{L,(i,s)}(t_2)|^2dP\\ &\le D\epsilon^2.\end{aligned}$$

This indicates that the bracketing number of $\mathscr{H}_\mathscr{F}$ associated with $L_2(P)$ norm, denoted by $N_{[\ \ ]}(\epsilon,\mathscr{H}_\mathscr{F},L_2(P))$, is bounded by $N_1 \times N_2$. It follows that $\log\left(N_{[\ \ ]}(\epsilon,\mathscr{H}_\mathscr{F},L_2(P))\right) \le \log(N_1 \times N_2) \le D/\epsilon$ for some constant D. Hence,

$$\int_0^1\sqrt{\log N_{[\ \ ]}(\epsilon,\mathscr{H}_\mathscr{F},L_2(P))}d\epsilon \le \int_0^1 D\epsilon^{-1/2}d\epsilon<\infty.$$

By Theorem 19.5 of van der Vaart and Wellner (1996, p. 270), $\mathscr{H}_\mathscr{F}$ is a P-Donsker Class.

Proof of Theorem 1 Let

$$\begin{aligned}\bar{\mathscr{L}}_n\left(\alpha,\widehat{S}_1,\widehat{S}_2;\mathbf{X}\right) &= \frac{1}{n}\sum_{i=1}^n l\left(\alpha,\widehat{S}_1(t_{1i}),\widehat{S}_2(t_{2i}),\delta_{1i},\delta_{2i}\right)\\ \bar{\mathscr{L}}_n(\alpha,S_1,S_2;\mathbf{X}) &= \frac{1}{n}\sum_{i=1}^n l(\alpha,S_1(t_{1i}),S_2(t_{2i}),\delta_{1i},\delta_{2i}),\end{aligned}$$

where $l(\alpha, \hat{S}_1(t_{1i}), \hat{S}_2(t_{2i})$ is defined in (3). First we show that $\bar{\mathscr{L}}_n\left(\alpha,\widehat{S}_1,\widehat{S}_2;\mathbf{X}\right) \xrightarrow{p} \mathrm{E}_{\alpha_0} l(\alpha,S_1(T_1),S_2(T_2),\Delta_1,\Delta_2)$ for any $\alpha \in A$.

By Taylor series expansion, we have

$$\begin{aligned}\bar{\mathscr{L}}_n\left(\alpha,\widehat{S}_1,\widehat{S}_2;\mathbf{X}\right)&=\bar{\mathscr{L}}_n\left(\alpha,S_1,S_2;\mathbf{X}\right)\\&+\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{S}_1\left(t_{1i}\right)-S_1\left(t_{1i}\right)\right)V_{\alpha,1}\left(\alpha,\tilde{S}_1\left(t_{1i}\right),\widehat{S}_2\left(t_{2i}\right),\delta_{1i},\delta_{2i}\right)\\&+\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{S}_2\left(t_{2i}\right)-S_2\left(t_{2i}\right)\right)V_{\alpha,2}\left(\alpha,\widehat{S}_1\left(t_{1i}\right),\tilde{S}_2\left(t_{2i}\right),\delta_{1i},\delta_{2i}\right)\end{aligned}$$

where $\sup_{t_1\in[0,t_{01}]}|\tilde{S}_1(t_1)-S_1(t_1)|\le\sup_{t_1\in[0,t_{01}]}|\widehat{S}_1(t_1)-S_1(t_1)|\xrightarrow{p}0$ (Fleming and Harrington 1991) and $\sup_{t_2\in[0,t_{02}]}|\tilde{S}_2(t_2)-S_2(t_2)|\le\sup_{t_2\in[0,t_{02}]}|\widehat{S}_2(t_2)-S_2(t_2)|\xrightarrow{p}0$ (Groeneboom and Wellner 1992), respectively.

By the Weak Law of Large Number Theorem,

$$\bar{\mathscr{L}}_n\left(\alpha,S_1,S_2;\mathbf{X}\right)\xrightarrow{p}\mathrm{E}_{\alpha_0}l\left(\alpha,S_1\left(T_1\right),S_2\left(T_2\right),\Delta_1,\Delta_2\right).$$

Note that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{S}_1\left(t_{1i}\right)-S_1\left(t_{1i}\right)\right)V_{\alpha,1}\left(\alpha,\tilde{S}_1\left(t_{1i}\right),\widehat{S}_2\left(t_{2i}\right),\delta_{1i},\delta_{2i}\right)\right|\le\sup_{t_1\in[0,t_{01}]}|\widehat{S}_1(t_1)-S_1(t_1)|\,\mathbb{P}_n|V_{\alpha,1}\left(\alpha,\tilde{S}_1\left(T_1\right),\widehat{S}_2\left(T_2\right),\Delta_1,\Delta_2\right)|$$

Denote $|\mathscr{G}_{\mathscr{F}}|=\left\{|V_{\alpha,1}(\alpha,f_1(t_1),f_2(t_2),\delta_1,\delta_2)|;f_j\in\mathscr{F}_j,\ j=1,2\right\}$. Since $\mathscr{G}_{\mathscr{F}}$ is a P-Glivenko-Cantelli class by Lemma 1, a straightforward algebra yields that the ε-bracketing number of $|\mathscr{G}_{\mathscr{F}}|$ is the same as the ε-bracketing number of $\mathscr{G}_{\mathscr{F}}$ which results in $|\mathscr{G}_{\mathscr{F}}|$ being a P-Glivenko-Cantelli class as well. Hence

$$\begin{aligned}\mathbb{P}_n|V_{\alpha,1}\left(\alpha,\tilde{S}_1\left(T_1\right),\widehat{S}_2\left(T_2\right),\Delta_1,\Delta_2\right)|&=P|V_{\alpha,1}\left(\alpha,\tilde{S}_1\left(T_1\right),\widehat{S}_2\left(T_2\right),\Delta_1,\Delta_2\right)|+o_p(1)\\&=P|V_{\alpha,1}\left(\alpha,S_1\left(T_1\right),S_2\left(T_2\right),\Delta_1,\Delta_2\right)|+o_p(1),\end{aligned}$$

due to the uniform consistency of $\tilde{S}_1(\cdot)$ and $\hat{S}_2(\cdot)$, the continuous mapping theorem, assumption (A2), and the dominated convergence theorem. This implies that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{S}_1\left(t_{1i}\right)-S_1\left(t_{1i}\right)\right)V_{\alpha,1}\left(\alpha,\tilde{S}_1\left(t_{1i}\right),\widehat{S}_2\left(t_{2i}\right),\delta_{1i},\delta_{2i}\right)\right|\xrightarrow{p}0.$$

Similar argument leads that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{S}_2\left(t_{2i}\right)-S_2\left(t_{2i}\right)\right)V_{\alpha,2}\left(\alpha,\widehat{S}_1\left(t_{1i}\right),\tilde{S}_2\left(t_{2i}\right),\delta_{1i},\delta_{2i}\right)\right|\xrightarrow{p}0.$$

This concludes $\bar{\mathscr{L}}_n\left(\alpha,\widehat{S}_1,\widehat{S}_2;\mathbf{X}\right)\xrightarrow{p}\mathrm{E}_{\alpha_0}l\left(\alpha,S_1\left(T_1\right),S_2\left(T_2\right),\Delta_1,\Delta_2\right)$. Now, $\forall\alpha\in A$, using Jensen's inequality, it follows that

$$\begin{aligned}&\bar{\mathscr{L}}_n\left(\alpha,\widehat{S}_1,\widehat{S}_2;\mathbf{X}\right)\\&\quad-\bar{\mathscr{L}}_n\left(\alpha_0,\widehat{S}_1,\widehat{S}_2;\mathbf{X}\right)\xrightarrow{p}\mathrm{E}_{\alpha_0}l\left(\alpha,S_1\left(T_1\right),S_2\left(T_2\right),\Delta_1,\Delta_2\right)\\&\quad-\mathrm{E}_{\alpha_0}l\left(\alpha_0,S_1\left(T_1\right),S_2\left(T_2\right),\Delta_1,\Delta_2\right)\\&\quad=\mathrm{E}_{\alpha_0}\log\frac{h\left(\alpha,T_1,T_2,\Delta_1,\Delta_2\right)}{h\left(\alpha_0,T_1,T_2,\Delta_1,\Delta_2\right)}<\log\mathrm{E}_{\alpha_0}\frac{h\left(\alpha,T_1,T_2,\Delta_1,\Delta_2\right)}{h\left(\alpha_0,T_1,T_2,\Delta_1,\Delta_2\right)}=0.\end{aligned}$$

Due to the convergence demonstrated above, $\forall\epsilon,\ \delta>0$, for which $(\alpha_0-\varepsilon,\ \alpha_0+\varepsilon)\in A$, we may find an integer $N=N(\varepsilon,\ \delta)$, such that, if $n>N$, for $\alpha=\alpha_0\pm\varepsilon$,

$$P\left(\bar{\mathscr{L}}_n\left(\alpha,\widehat{S}_1,\widehat{S}_2;\mathbf{X}\right)<\bar{\mathscr{L}}_n\left(\alpha_0,\widehat{S}_1,\widehat{S}_2;\mathbf{X}\right)\right)>1-\delta.$$

Thus for $n>N$,

$P\left(\bar{\mathscr{L}}_n\left(\alpha,\widehat{S}_1,\widehat{S}_2;\mathbf{X}\right)\right.$ has a local maximum $\left.\widehat{\alpha}_n\in(\alpha_0-\epsilon,\alpha_0+\epsilon)\right)>1-2\delta$, because of (A1). This immediately shows that the sequence of random variables $\widehat{\alpha}_n$ converge in probability to α_0 as $n\to\infty$.

Proof of Theorem 2 Under (A1), Taylor expansion of the pseudo score function gives

$$\begin{aligned}0=&\mathbb{P}_nV_\alpha\left(\widehat{\alpha}_n,\widehat{S}_1\left(T_1\right),\widehat{S}_2\left(T_2\right),\Delta_1,\Delta_2\right)\\&=\mathbb{P}_nV_\alpha\left(\alpha_0,\widehat{S}_1\left(T_1\right),\widehat{S}_2\left(T_2\right),\Delta_1,\Delta_2\right)\\&+\left(\widehat{\alpha}_n-\alpha_0\right)\mathbb{P}_nV_{\alpha^2}\left(\alpha_0,\widehat{S}_1\left(T_1\right),\widehat{S}_2\left(T_2\right),\Delta_1,\Delta_2\right)\\&+O_p\left(\left|\widehat{\alpha}_n-\alpha_0\right|^2\right),\end{aligned}$$

then we get

$$\sqrt{n}\left(\widehat{\alpha}_n-\alpha_0\right)=\frac{\sqrt{n}\mathbb{P}_nV_\alpha\left(\alpha_0,\widehat{S}_1\left(T_1\right),\widehat{S}_2\left(T_2\right),\Delta_1,\Delta_2\right)}{-\mathbb{P}_nV_{\alpha^2}\left(\alpha_0,\widehat{S}_1\left(T_1\right),\widehat{S}_2\left(T_2\right),\Delta_1,\Delta_2\right)-O_p\left(\left|\widehat{\alpha}_n-\alpha_0\right|\right)}.$$

First, we show that

$$\mathbb{P}_nV_{\alpha^2}\left(\alpha_0,\widehat{S}_1\left(T_1\right),\widehat{S}_2\left(T_2\right),\Delta_1,\Delta_2\right)\xrightarrow{p}W\left(\alpha_0,S_1,S_2\right),$$

where

$$\begin{aligned}W\left(\alpha_0,S_1,S_2\right)&=PV_{\alpha^2}\left(\alpha_0,S_1\left(T_1\right),S_2\left(T_1\right),\Delta_1,\Delta_2\right)\\&=-P[V_\alpha\left(\alpha_0,S_1\left(T_1\right),S_2\left(T_2\right),\Delta_1,\Delta_2\right)]^2.\end{aligned}$$

We can rewrite $\mathbb{P}_n V_{\alpha^2}\left(\alpha_0, \widehat{S}_1(T_1), \widehat{S}_2(T_2), \Delta_1, \Delta_2\right) = \mathbb{P}_n V_{\alpha^2}(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2) + R_n$. By the uniform consistency of $\hat{S}_1$ and $\hat{S}_2$ and the fact that $V_{\alpha^2}(\alpha_0, S_1(t_1), S_2(t_2), \delta_1, \delta_2)$ satisfies the Lipschitz condition due to (A2), it follows that

$$\mathbb{P}_n V_{\alpha^2}\left(\alpha_0, \widehat{S}_1(T_1), \widehat{S}_2(T_2), \Delta_1, \Delta_2\right) = \mathbb{P}_n V_{\alpha^2}(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2) + o_p(1).$$

This results in

$$\mathbb{P}_n V_{\alpha^2}\left(\alpha_0, \widehat{S}_1(T_1), \widehat{S}_2(T_2), \Delta_1, \Delta_2\right) \xrightarrow{p} PV_{\alpha^2}(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2)$$

by the Weak Law of Large Number Theorem.

Second, we derive the asymptotic distribution of $\sqrt{n}\mathbb{P}_n V_\alpha\left(\alpha_0, \widehat{S}_1(T_1), \widehat{S}_2(T_2), \Delta_1, \Delta_2\right)$. Note that

$$\begin{aligned}\mathbb{P}_n V_\alpha\left(\alpha_0, \widehat{S}_1(T_1), \widehat{S}_2(T_2), \Delta_1, \Delta_2\right) &= (\mathbb{P}_n - P)\left(V_\alpha\left(\alpha_0, \widehat{S}_1(T_1), \widehat{S}_2(T_2), \Delta_1, \Delta_2\right) - V_\alpha(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2)\right) + \mathbb{P}_n V_\alpha(\alpha_0, S_1(T_1) \\ &= u_{1n} + u_{2n} + u_{3n}.\end{aligned}$$

Lemma 2 indicates that under (A1) and (A2), $\mathscr{H}_{\mathscr{F}}$ is a P-Donsker class. Furthermore, since $\sup_{0 \le t_j \le t_{0j}} |\widehat{S}_j(t_j) - S_j(t_j)| \xrightarrow{p} 0$, $j=1, 2$, by the Dominated Convergence Theorem,

$$\int \left(\widehat{S}_j(t_j) - S_j(t_j)\right)^2 dP(t_1, t_2, \delta_1, \delta_2) \xrightarrow{p} 0,\ j=1, 2.$$

Therefore, $\sqrt{n}u_{1n} = o_p(1)$ by Lemma 19.24 of van der Vaart and Wellner (1996).

Note that u_{2n} is a sum of independent and identically distributed quantities, where each quantity has mean

$$\int V_\alpha(\alpha_0, S_1(t_1), S_2(t_2), \delta_1, \delta_2)\, dP(t_1, t_2, \delta_1, \delta_2) = 0$$

and variance

$$\int [V_\alpha(\alpha_0, S_1(t_1), S_2(t_2), \delta_1, \delta_2)]^2 dP(t_1, t_2, \delta_1, \delta_2) = -W(\alpha_0, S_1, S_2).$$

By the Central Limit Theorem, $\sqrt{n}u_{2n}$ converges in distribution to a normal random variable with mean 0 and variance $-W(\alpha_0, S_1, S_2)$.

Applying Von Mises Expansion (von Mises 1947) on u_{3n} around S_1, S_2, we get

$$u_{3n} \overset{d}{=} \int_0^{t_{01}} IC_1(t_1)\, d\left(\widehat{S}_1 - S_1\right)(t_1) + \int_0^{t_{02}} IC_2(t_2)\, d\left(\widehat{S}_2 - S_2\right)(t_2), \quad (4)$$

where $IC_j(t)$, $j = 1, 2$ are the influence curves of the functional $PV_\alpha(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2)$ which are defined by

$$\begin{aligned} IC_1(t_1) &= -\int_0^{t_1}\int_0^{t_{02}} V_{\alpha,1}(\alpha_0, S_1(\tau_1), S_2(\tau_2), \delta_1, \delta_2)\, dP(\tau_1, \tau_2, \delta_1, \delta_2) \\ &= \int_0^{t_1}\int_0^{t_{02}} M_{\alpha,1}(\alpha_0, S_1(\tau_1), S_2(\tau_2))\, f(\tau_1, \tau_2)\, d\tau_1 d\tau_2 \end{aligned}$$

and

$$\begin{aligned} IC_2(t_2) &= -\int_0^{t_2}\int_0^{t_{01}} V_{\alpha,2}(\alpha_0, S_1(\tau_1), S_2(\tau_2), \delta_1, \delta_2)\, dP(\tau_1, \tau_2, \delta_1, \delta_2) \\ &= \int_0^{t_2}\int_0^{t_{01}} M_{\alpha,2}(\alpha_0, S_1(\tau_1), S_2(\tau_2))\, f(\tau_1, \tau_2)\, d\tau_1 d\tau_2, \end{aligned} \quad (5)$$

respectively. Here

$$\begin{aligned} M_{\alpha,1}(\alpha_0, S_1(\tau_1), S_2(\tau_2)) &= -\mathrm{E}\{V_{\alpha,1}(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2) | T_1=\tau_1, T_2=\tau_2\} \\ M_{\alpha,2}(\alpha_0, S_1(\tau_1), S_2(\tau_2)) &= -\mathrm{E}\{V_{\alpha,2}(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2) | T_1=\tau_1, T_2=\tau_2\}. \end{aligned}$$

Using the martingale theory for counting process, Pepe (1991) showed that, for $t \in [0, t_{01}]$, $(\hat{S}_1(t_1) - S_1(t_1))$ is asymptotically equivalent to a sum of n i.i.d. random variables $\sum_i I_1^0(T_{1i}, \Delta_{1i})(t_1)/n$. It follows that

$$\int_0^{t_{01}} IC_1(t_1)\, d\left(\widehat{S}_1 - S_1\right)(t_1) = \frac{1}{n}\sum_{i=1}^{n} I_1(T_{1i}, \Delta_{1i}; \alpha_0) + o_p(1), \quad (6)$$

where

$$I_1(T_{1i}, \Delta_{1i}; \alpha_0) = \int_0^{t_{01}}\int_0^{t_{02}} M_{\alpha,1}(\alpha_0, S_1(\tau_1), S_2(\tau_2))\, f(\tau_1, \tau_2)\, I_1^0(T_{1i}, \Delta_{1i})(\tau_1)\, d\tau_1 d\tau_2$$

and I_1^0 is a martingale given by

$$I_1^0(T_1, \Delta_1)(t_1) = -S_1(t_1)\left\{\int_0^{t_1} \frac{1}{P(T_1 \geq u)} dN_1(u) - \int_0^{t_1} \frac{I[T_1 \geq u]}{P(T_1 \geq u)} d\Lambda_1(u)\right\},$$

in which $N_1(u)$ is defined as $I[T_1 \leq u, \Delta_1 = 1]$ and Λ_1 is the cumulative hazard function of T_1^0.

On the other hand, although $(\hat{S}_2 - S_2)(t_2)$ can not be written as sum of i.i.d random quantities, a smooth functional of the nonparametric maximum likelihood estimator $\hat{S}_2$ can still be shown asymptotically normal (Huang and Wellner 1995). Using this property and the regularity conditions (A3)–(A6), Wang and Ding (2000) showed that

$$\int_0^{t_{02}} IC_2(t_2)\, d\left(\widehat{S}_2 - S_2\right)(t_2) = -\mathbb{P}_n \tilde{l}(\,_2, \Delta_2; S_2, G_2, \psi_2) + o_p(1) \quad (7)$$

with $\tilde{l}(T_2, \Delta_2; S_2, G_2, \psi_2) = -[\Delta_2 - (1 - S_2(T_2))]\frac{\psi_2(T_2)}{g_2(T_2)} I[g_2(T_2) > 0]$ and thus $\sqrt{n}\int_0^{t_{02}} IC_2(t_2)\, d\left(\widehat{S}_2 - S_2\right)(t_2)$ converges in distribution to a normal random variable with mean 0.

In summary, we obtain that,

$$\begin{aligned}\mathbb{P}_n V_\alpha\left(\alpha_0, \widehat{S}_1(T_1), \widehat{S}_2(T_2), \Delta_1, \Delta_2\right) &= \mathbb{P}_n\left[V_\alpha(\alpha_0, S_1(T_1), S_2(T_2), \Delta_1, \Delta_2) + I_1(T_1, \Delta_1; \alpha_0) - \tilde{l}(T_2, \Delta_2; S_2, G_2, \psi_2)\right] + o_p\left(n^{-1/2}\right) \\ &= \mathbb{P}_n Q(T_1, T_2, \Delta_1, \Delta_2; \alpha_0, S_1, S_2) + o_p\left(n^{-1/2}\right).\end{aligned}$$

Therefore, $\sqrt{n}\mathbb{P}_n V_\alpha\left(\alpha_0, \widehat{S}_1(T_1), \widehat{S}_2(T_2), \Delta_1, \Delta_2\right)$ is asymptotically normal with mean zero and variance Var($Q(T_1, T_2, \Delta_1, \Delta_2; \alpha_0, S_1, S_2)$). Hence,

$$\sqrt{n}(\widehat{\alpha}_n - \alpha_0) \xrightarrow{d} N\left(0, \sigma^2\right),$$

where

$$\sigma^2 = \frac{\mathrm{Var}\left(Q(T_1, T_2, \Delta_1, \Delta_2; \alpha_0, S_1, S_2)\right)}{W^2(\alpha_0, S_1, S_2)}.$$

References

Birk M, Lindback S, Lidman C. No influence of GB virus C replication on the prognosis in a cohort of HIV-1-infected patients. AIDS. 2002; 16:2482–2485. [PubMed: 12461426]

Ding AA, Wang W. Testing independence for bivariate current status data. J Am Stat Assoc. 2004; 99:145–155.

Fleming, TR.; Harrington, DP. Counting process and survival analysis. John wiley & Sons; New York: 1991.

Groeneboom, P.; Wellner, JA. Information bounds and nonparametric maximum likelihood estimation. Birkhauser; Boston: 1992.

Hougaard P. Fitting a multivariate failure time distribution. IEEE Trans Reliab. 1989; 38:444–448.

Hsu L, Prentice RL. A generalisation of the mantel-haenszel test to bivariate failure time data. Biometrika. 1996; 4:905–911.

Huang J, Wellner JA. Asymptotic normality of the npmle of linear functionals for interval censored data, case 1. Stat Neerl. 1995; 49:153–163.

Kalbfleisch, JD.; Prentice, RL. The statistical analysis of failure time data. 2nd edn.. Wiley-Interscience; New York: 2002.

Liang KE, Self SG, Bandeen-Rocche K, Zeger S. Some recent developments for regression analysis of multivariate failure time data. Lifetime Data Anal. 1995; 1:403–415. [PubMed: 9385112]

Nelsen, RB. An introduction to copulas. 2nd edn.. Springer-Verlag; New York: 2006.

Oakes D. A concordance test for independence in the presence of censoring. Biometrics. 1982; 38:451–455. [PubMed: 7052151]

Oakes D. Bivariate survival models induced by frailties. J Am Stat Assoc. 1989; 84:487–493.

Oakes D. Survival analysis. J Am Stat Assoc. 2000; 95:282–285.

Pepe MS. Inference for events with dependent risks in multiple endpoint studies. J Am Stat Assoc. 1991; 86:770–778.

Shih JH, Louis T. Inference on the association parameter in copula models for bivariate survival data. Biometrics. 1995; 51:1384–1399. [PubMed: 8589230]

Shih JH, Louis TA. Tests of independence for bivariate survival data. Biometrics. 1996; 4:1440–1449. [PubMed: 8962462]

Tillmann H, Heiken H, Knapik-Botor A, Heringlake S, Ockenga J, et al. Infection with GB virus C and reduced mortality among HIV-infected patients. N Engl J Med. 2001; 345:715–724. [PubMed: 11547740]

Toyoda H, Fukuda Y, et al. Effect of GB virus C/hepatitis G virus coinfection on the course of HIV infection in hemophilia patients in Japan. J Acquir Immune Defic Syndr Hum Retrovirol. 1998; 17:209–213. [PubMed: 9495219]

van der Vaart, AW. Asymptotic statistics. Cambridge Univ. Press; Cambridge: 1998.

van der Vaart, AW.; van der Wellner, JA. Weak convergence and empirical processes with application to statistics. Springer-Verlag; New York: 1996.

von Mises R. On the asymptotic distribution of differentiable statistical functions. Ann Math Statist. 1947; 18:309–348.

Wang W, Ding AA. On assessing the association for bivariate current status data. Biometrika. 2000; 87:879–893.

Williams C, Klinzman D, Yamashita T, Xiang J, et al. Persistent GB virus C infection and survival in HIV-infected men. N Engl J Med. 2004; 350:981–990. [PubMed: 14999110]

Xiang J, Wunschmann W, Diekema D, Klinzman D, Patrick K, et al. Effect of coinfection with GB virus C on survival among patients with HIV infection. N Engl J Med. 2001; 345:707–714. [PubMed: 11547739]

Zhang W, Chaloner K, Tillmann HS, Williams CF, Stapleton JT. Effect of early and late GBV-C viremia on survival of HIV-infected individuals: a meta-analysis. HIV Med. 2006; 7:173–180. [PubMed: 16494631]

Table 1

Simulation results of the two-stage maximum pseudo-likelihood estimator based on 1,000 Monte-Carlo samples with sample size ranged from 50 to 400 for $\alpha = 4/3, 2, 4$

		$n = 50$		$n = 100$		$n = 200$		$n = 400$	
		$\widehat{\alpha}_n$	$\widehat{\tau}_n$	$\widehat{\alpha}_n$	$\widehat{\tau}_n$	$\widehat{\alpha}_n$	$\widehat{\tau}_n$	$\widehat{\alpha}_n$	$\widehat{\tau}_n$
$\tau = 0.25$	Bias	0.219	0.043	0.059	0.013	0.023	0.005	–0.005	–0.002
$\alpha = 1.333$	ese	0.845	0.172	0.233	0.113	0.142	0.076	0.098	0.055
	bse	8.799	0.161	0.334	0.109	0.154	0.077	0.099	0.054
	95% ecp	0.968		0.966		0.963		0.954	
$\tau = 0.50$	Bias	1.120	0.051	0.194	0.021	0.102	0.014	0.032	0.003
$\alpha = 2.0$	ese	8.090	0.158	0.563	0.098	0.320	0.070	0.208	0.050
	bse	26.276	0.156	2.523	0.101	0.359	0.069	0.213	0.048
	95% ecp	0.985		0.976		0.966		0.957	
$\tau = 0.75$	Bias	9.460	0.176	0.695	0.037	0.189	0.017	0.058	0.004
$\alpha = 4.0$	ese	51.64	0.117	4.305	0.081	1.002	0.054	0.646	0.038
	bse	60.48	0.124	15.726	0.079	1.597	0.054	0.696	0.038
	95% ecp	0.991		0.980		0.974		0.959	

Table 2

Simulation results of maximum likelihood analysis (S_1 and S_2 are known) based on 1,000 Monte-Carlo samples with sample size ranged from 50 to 400 for α = 4/3, 2, 4

		***n* = 50**		***n* = 100**		***n* = 200**		***n* = 400**	
		$\tilde{\alpha}_n$	$\tilde{\tau}_n$	$\tilde{\alpha}_n$	$\tilde{\tau}_n$	$\tilde{\alpha}_n$	$\tilde{\tau}_n$	$\tilde{\alpha}_n$	$\tilde{\tau}_n$
τ = 0.25	Bias	0.065	0.031	0.019	0.011	0.016	0.004	–0.004	–0.001
α = 1.333	ese	0.334	0.145	0.192	0.102	0.136	0.076	0.097	0.053
	bse	1.253	0.136	0.218	0.101	0.136	0.073	0.094	0.053
	95% ecp	0.940		0.942		0.954		0.949	
τ = 0.50	Bias	0.302	0.036	0.069	0.018	0.022	0.005	–0.009	–0.002
α = 2.0	ese	1.360	0.141	0.455	0.096	0.288	0.068	0.190	0.047
	bse	8.808	0.132	0.811	0.093	0.288	0.068	0.196	0.047
	95% ecp	0.965		0.951		0.939		0.952	
τ = 0.75	Bias	7.136	0.160	0.539	0.030	0.164	0.010	0.014	0.001
α = 4.0	ese	41.24	0.104	3.830	0.073	0.975	0.050	0.635	0.038
	bse	44.55	0.089	12.353	0.067	1.590	0.049	0.659	0.036
	95% ecp	0.971		0.963		0.960		0.940	

Table 3

The analysis of association between HIV survival time and GBV-C persistence time: include all subjects whose GBV-C at early visit are positive ($N = 107$)

	Estimate	Bootstrap SE	95% Wald CI	95% Bootstrap CI
$\widehat{\alpha}_n$	1.5836	0.2037	[1.1843, 1.9829]	[1.2043, 2.0598]
$\widehat{\tau}_n$	0.3685	0.0866	[0.1988, 0.5383]	[0.2114, 0.5533]