Statistica Sinica Preprint No: SS-2019-0296	
Title	A spline-based nonparametric analysis for interval-censored bivariate survival data
Manuscript ID	SS-2019-0296
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0296
Complete List of Authors	Yuan Wu, Ying Zhang and Junyi Zhou
Corresponding Author	Yuan Wu
E-mail	yuan.wu@duke.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

A spline-based nonparametric analysis for interval-censored bivariate survival data

Yuan Wu

Duke University Medical Center, Durham, NC 27705

Ying Zhang

University of Nebraska Medical Center, Omaha, NE 68198

Junyi Zhou

Indiana University Fairbanks School of Public Health, Indianapolis, IN 46202

Abstract: In this manuscript we propose a spline-based sieve nonparametric maximum likelihood estimation method for joint distribution function with bivariate interval-censored data. We study the asymptotic behavior of the proposed estimator by proving the consistency and deriving the rate of convergence. Based on the sieve estimate of the joint distribution, we also develop an efficient nonparametric test for making inference about the dependence between two interval-censored event times and establish its asymptotic normality. We conduct simulation studies to examine the finite sample performance of the proposed methodology. Finally we apply the method to assess the association between two subtypes

of mild cognitive impairment (MCI): amnestic MCI and non-amnestic MCI, for Huntington disease (HD) using data from a 12-year observational cohort study on premanifest HD individuals, PREDICT-HD.

Key words and phrases: Empirical process, Generalized gradient projection algorithm, Sieve Estimation.

1. Introduction

Interval-censored time-to-event data occur very often in clinical and other biomedical studies. Interval censoring means that one only knows the event time of interest lies in a time interval that is normally derived from consecutive observation time points. A special case for interval censoring is called current status data, for which only left or right censoring happens, that is, either the left end point of each observation interval is zero or the right end point of that observation interval is infinity. The importance of studying interval-censored time-to-event data has been well recognized. Research on statistical inference for interval-censored data has been an active area in nonparametric or semiparametric statistical modeling, which includes Turnbull (1976) and Groeneboom and Wellner (1992) for nonparametric maximum likelihood estimation (NPMLE); Sun (1996), Fay (1999) and Zhang et al. (2001) for comparing survival functions among different

exposure groups; and Huang (1996), Zhang et al. (2010) and Wang et al. (2016) for semiparametric regression analysis.

Research of multivariate interval-censored time-to-event data, in particular bivariate interval-censored data has been an important but very challenging topic in statistical literatures. Betensky and Finkelstein (1999) and Wong and Yu (1999) are among the several earliest papers to study the conventional NPMLE of the joint survival function with bivariate interval-censored data. Maathuis (2005) proposed a fast and stable algorithm to compute the MPMLE. In an unpublished dissertation, Song (2001) described the consistency and convergence rate of the NPMLE with bivariate interval-censored data. Despite these efforts, the conventional NPMLE of the joint survival function with interval-censored data has been known not uniquely determined and its asymptotic behavior has not been completely justified. Wu and Zhang (2012) proposed a spline-based sieve NPMLE of the joint distribution function with bivariate current status data. Under some mild regularity conditions, they proved the consistence and derived the rate of convergence that is better than the rate given by Song (2001) for the conventional NPMLE.

Semiparametric regression analyses under Copula or frailty model were adopted for bivariate interval-censored data in recent literature. Wen and

Chen (2013) proposed to use a frailty model approach for the semiparametric bivariate interval-censored data with the marginal hazards for time-to-event data modelled by the Cox proportional hazards (PH) model. Zeng et al. (2017) and Zhou et al. (2017) both extended Wen and Chen (2013)'s work to allow more general semiparametric models for the marginal time-to-event data: Zeng et al. (2017) included random effects for the covariates; Zhou et al. (2017) adopted the spline-based sieve estimation. Hu et al. (2017) proposed to use a Copula model for analyzing bivariate current status data. Specifically they used the Bernstein polynomial-based copula to construct the joint distribution function along with the Cox PH model for marginal time-to-event data.

Testing for the dependence between bivariate time-to-event data has been a common practice in statistical applications. It is particularly important in biomedical research that people may experience two or more adverse clinical events and understanding of the associations between the events will help enhance the study of risk factors for the events. For example, if epidemiologists want to study the risk factors for a rare disease and know this disease is strongly associated with another clinical event, which is more commonly observed, choosing this event as the surrogate endpoint of the rare disease to ascertain the risk factors may largely improve the study effi-

ciency. Wang and Ding (2000) adopted the idea for bivariate right-censored data by Shih and Louis (1995) and proposed a two-stage approach to test the association parameter based on a Copula model with bivariate current status data. Specially, in the first stage Wang and Ding (2000) computed the conventional NPMLEs for both marginal distributions of event times and in the second stage they developed pseudo-MLE method for the association parameter by plugging the NPMLEs from the first stage to the likelihood based on the bivariate copula model. Sun et al. (2006) extended the Wang-Ding's method to analyze bivariate interval-censored data. Following the same idea by Shih and Louis (1996) for testing the association of the two event times under the right censoring mechanism, Ding and Wang (2004) developed a nonparametric test for the independence between two event times with bivariate current status data, which can be viewed as a generalization of Mantel-Haenszel test. Jewell et al. (2005) considered a special case of bivariate current status data, in which the observation time for both events is the same. For this case, the dependence test statistic can be constructed based on a functional of the NPMLEs for the marginal distribution functions. Kim et al. (2015) adopted the approach for bivariate right censoring by Brown et al. (1974) and developed an association test based on estimating $Kendall's\ \tau$ for bivariate interval-censored data.

However, the asymptotic normality of the test statistic given by Kim et al. (2015) seems difficult to justify theoretically when the majority of censoring rectangles are overlapped. It is noted that all aforementioned inferences for interval-censored data were either under specific model structure for the joint distribution (Wen and Chen, 2013; Zeng *et al.*, 2017; Zhou *et al.*, 2017; Hu *et al.*, 2017; Wang and Ding, 2000; Sun *et al.*, 2006), or to deal with special cases of interval-censored data (Ding and Wang, 2004; Jewell *et al.*, 2005) or was quite ad-hoc (Kim et al., 2015). To the best of our knowledge there is no rigorously justified model-free nonparametric method in the literature for the inference of association between bivariate interval-censored time-to-event data.

This work is motivated by a 12-year internationally multi-sites observational study of premanifest Huntington disease (HD) patients to identify the neurobiological predictors of HD onset, PREDICT-HD. A predominantly motor impaired neurodegenerative disease, HD also leads to cognitive impairments, possibly in multiple domains. As an early sign of disease progression, mild cognitive impairment (MCI) is commonly studied in neurodegenerative diseases and may be chosen as a study endpoint for clinical trials to treat HD patients. There is a great interest in the HD research community to study the age of MCI onset in premanifest HD patients and correlations

among the ages of onset in different subtypes of MCI. For the detailed description of the PREDICT-HD study, we defer to the section of real data analysis. In this paper, we aim to develop a spline-based sieve nonparametric maximum likelihood estimation method for bivariate interval-censored data and to construct an efficient statistical test for the association between two event times under bivariate interval-censored data model. The test is based on a functional of the sieve NPMLEs of the joint and marginal distribution functions. We apply the proposed method to the PREDICT-HD data.

The reminder of this paper is organized as follows. Section 2 proposes the spline-based sieve NPMLEs for the distribution functions and constructs a test statistic for the association. Section 3 establishes three theorems to describe the asymptotic behavior of the spline-based sieve NPMLEs and the asymptotic normality of the proposed test statistic. Section 4 outlines the algorithm to compute the sieve NPMLEs. Section 5 conducts simulation studies to justify the sieve NPMLE of the joint distribution function and the test statistic. Section 6 applies the proposed method to the PREDICT-HD to ascertain the possible association of ages of onset in two MCI subtypes: amnestic MCI and non-amnestic MCI. Finally, Section 7 discusses some existing issues and closely related future work. In addition, the technical

details including lemmas and their proofs are presented in the supplementary material.

2. Method

We first propose a spline-based sieve NPMLE method for the joint and marginal distribution functions and then develop a nonparametric association test for the dependence between the bivariate event times based on a functional of the sieve NPMLEs.

2.1 Spline-based Sieve NPMLEs for the Distribution Functions

Assume bivariate event times T_1 and T_2 are interval censored by (U_1, V_1) and (U_2, V_2), respectively. Suppose a sample of size n for censoring times with their relationship to event times is given by

$$\left[\left\{u_{1,k}, v_{1,k}, u_{2,k}, v_{2,k}, \left(\delta_{1,k}^{(j)}, \delta_{2,k}^{(j)}\right)_{j=1}^{3}\right\}\right]_{k=1}^{n},$$

where $\{(u_{1,k}, v_{1,k}, u_{2,k}, v_{2,k})\}_{k=1}^{n}$ is the sample for (U_1, V_1, U_2, V_2); $\delta_{i,k}^{(1)} = 1_{[t_{i,k} \le u_{i,k}]}$, $\delta_{i,k}^{(2)} = 1_{[u_{i,k} < t_{i,k} \le v_{i,k}]}$ and $\delta_{i,k}^{(3)} = 1_{[t_{i,k} > v_{i,k}]}$ respectively indicate left censoring, interval censoring and right censoring, and $\{t_{i,k}\}_{k=1}^{n}$, the sample of unobserved T_i for $i = 1, 2$. Suppose that event times are independent to censoring times. Let $\boldsymbol{\theta} = (F_0(\cdot,\cdot), F_1(\cdot), F_2(\cdot))$ with F_0, F_1 and F_2, respectively, denoting the joint distribution function of event times T_1 and T_2, the

marginal distribution functions of T_1 and T_2. Then the log likelihood of the model parameter $\boldsymbol{\theta}$ based on the n observations can be written as

$$
\begin{aligned}
&l_n(\boldsymbol{\theta};\text{data}) \\
&= \sum_{k=1}^{n} \{ \delta_{1,k}^{(1)} \delta_{2,k}^{(1)} \log F_0(u_{1,k}, u_{2,k}) \\
&\quad + \delta_{1,k}^{(1)} \delta_{2,k}^{(2)} \log[F_0(u_{1,k}, v_{2,k}) - F_0(u_{1,k}, u_{2,k})] \\
&\quad + \delta_{1,k}^{(1)} \delta_{2,k}^{(3)} \log[F_1(u_{1,k}) - F_0(u_{1,k}, v_{2,k})] \\
&\quad + \delta_{1,k}^{(2)} \delta_{2,k}^{(1)} \log[F_0(v_{1,k}, u_{2,k}) - F_0(u_{1,k}, u_{2,k})] \\
&\quad + \delta_{1,k}^{(2)} \delta_{2,k}^{(2)} \log[F_0(v_{1,k}, v_{2,k}) - F_0(u_{1,k}, v_{2,k}) - F_0(v_{1,k}, u_{2,k}) + F_0(u_{1,k}, u_{2,k})] \\
&\quad + \delta_{1,k}^{(2)} \delta_{2,k}^{(3)} \log[F_1(v_{1,k}) - F_0(v_{1,k}, v_{2,k}) - F_1(u_{1,k}) + F_0(u_{1,k}, v_{2,k})] \\
&\quad + \delta_{1,k}^{(3)} \delta_{2,k}^{(1)} \log[F_2(u_{2,k}) - F_0(v_{1,k}, u_{2,k})] \\
&\quad + \delta_{1,k}^{(3)} \delta_{2,k}^{(2)} \log[F_2(v_{2,k}) - F_2(u_{2,k}) - F_0(v_{1,k}, v_{2,k}) + F_0(v_{1,k}, u_{2,k})] \\
&\quad + \delta_{1,k}^{(3)} \delta_{2,k}^{(3)} \log[1 - F_1(v_{1,k}) - F_2(v_{2,k}) + F_0(v_{1,k}, v_{2,k})] \}.
\end{aligned}
\tag{2.1}
$$

Please refer Section S1 of the online supplementary material for detailed derivation of the log likelihood. The conventional NPMLE method for estimating $\boldsymbol{\theta}$ is a challenging task both computationally and theoretically. We propose to adopt the spline-based sieve NPMLE method as originally proposed by Wu and Zhang (2012) for bivariate current status data to estimate $\boldsymbol{\theta}$ nonparametrically for (2.1). Suppose $T_1 \in [0, \tau_1]$ and $T_2 \in [0, \tau_2]$.

Denote two sets of B-splines basis functions of order l (Schumaker, 1981): $\{B_i^{(1),l}(t)\}_{i=1}^{p_n}$ with knot sequence $\boldsymbol{\xi}$ as

$$\begin{aligned}\boldsymbol{\xi} = \{&(\xi_i)_{i=1}^{p_n+l} : \\ &0 = \xi_1 = \cdots = \xi_l < \xi_{l+1} < \cdots < \xi_{p_n} < \xi_{p_n+1} = \xi_{p_n+l} = \tau_1\},\end{aligned} \tag{2.2}$$

and $\{B_j^{(2),l}(t)\}_{j=1}^{q_n}$ with the knot sequence $\boldsymbol{\eta}$ as

$$\begin{aligned}\boldsymbol{\eta} = \{&(\eta_j)_{j=1}^{q_n+l} : \\ &0 = \eta_1 = \cdots = \eta_l < \eta_{l+1} < \cdots < \eta_{q_n} < \eta_{q_n+1} = \eta_{q_n+l} = \tau_2\},\end{aligned} \tag{2.3}$$

where p_n and q_n are both positive integers related to n.

Let

$$F_{n,0}(\cdot,\cdot) = \sum_{i=1}^{p_n}\sum_{j=1}^{q_n} \alpha_{i,j} B_i^{(1),l}(\cdot) B_j^{(2),l}(\cdot), \tag{2.4}$$

$$F_{n,1}(\cdot) = \sum_{i=1}^{p_n} \beta_i B_i^{(1),l}(\cdot) \tag{2.5}$$

and

$$F_{n,2}(\cdot) = \sum_{j=1}^{q_n} \gamma_j B_j^{(2),l}(\cdot) \tag{2.6}$$

be the B-spline-based joint and marginal distribution functions, correspondingly (Bollaerts et al., 2006). First, we need to ensure that these functions satisfy the requirements for being the distribution functions as discussed in Wu and Zhang (2012). For $\boldsymbol{\theta}_n = (F_{n,0}, F_{n,1}, F_{n,2})$, we also need to ensure

that $l_n(\boldsymbol{\theta}_n; \text{data})$ is bounded for the existence of sieve NPMLEs. So we assume that there exist $\tau_{1,l} > 0$, $\tau_{1,h} < \tau_1$, $\tau_{2,l} > 0$, $\tau_{2,h} < \tau_2$ and $\tau_d > 0$, such that the domain for censoring times (U_1, V_1, U_2, V_2) is given by

$$\begin{aligned}\mathcal{D} = \{(u_1, v_1, u_2, v_2) : u_1 \in [\tau_{1,l}, \tau_{1,h}], v_1 \in [\tau_{1,l}, \tau_{1,h}], \\ u_2 \in [\tau_{2,l}, \tau_{2,h}], v_2 \in [\tau_{2,l}, \tau_{2,h}], u_1 + \tau_d \le v_1, u_2 + \tau_d \le v_2\}\end{aligned} \tag{2.7}$$

and for $(u_1, v_1, u_2, v_2) \in \mathcal{D}$, the following constraints are imposed for (2.4), (2.5) and (2.6).

$$\begin{aligned}
&0 < F_{n,0}(u_1, u_2), \\
&F_{n,0}(u_1, u_2) < F_{n,0}(v_1, u_2), \\
&F_{n,0}(u_1, u_2) < F_{n,0}(u_1, v_2), \\
&\{F_{n,0}(v_1, v_2) - F_{n,0}(u_1, v_2)\} - \{F_{n,0}(v_1, u_2) - F_{n,0}(u_1, u_2)\} > 0, \\
&F_{n,1}(u_1) - F_{n,0}(u_1, v_2) > 0, \\
&F_{n,2}(u_2) - F_{n,0}(v_1, u_2) > 0, \\
&\{F_{n,1}(v_1) - F_{n,1}(u_1)\} - \{F_{n,0}(v_1, v_2) - F_{n,0}(u_1, v_2)\} > 0, \\
&\{F_{n,2}(v_2) - F_{n,2}(u_2)\} - \{F_{n,0}(v_1, v_2) - F_{n,0}(v_1, u_2)\} > 0, \\
&\{1 - F_{n,1}(v_1)\} - \{F_{n,2}(v_2) - F_{n,0}(v_1, v_2)\} > 0.
\end{aligned} \tag{2.8}$$

Now we define the parameter space for spline-based distribution func-

tions as

$$\Psi_n = \Bigg\{ \boldsymbol{\theta}_n = (F_{n,0}, F_{n,1}, F_{n,2}) : F_{n,0}(\cdot,\cdot) = \sum_{i=1}^{p_n}\sum_{j=1}^{q_n} \alpha_{i,j} B_i^{(1),l}(\cdot) B_j^{(2),l}(\cdot),$$
$$F_{n,1}(\cdot) = \sum_{i=1}^{p_n} \beta_i B_i^{(1),l}(\cdot), F_{n,2}(\cdot) = \sum_{j=1}^{q_n} \gamma_j B_j^{(2),l}(\cdot),$$
$$(2.8) \text{ holds for } (u_1, v_1, u_2, v_2) \in \mathcal{D} \text{ with } \mathcal{D} \text{ defined by (2.7)},$$
$$\text{knot sequences are as (2.2) and (2.3)} \Bigg\}. \tag{2.9}$$

Then the proposed spline-based Sieve NPMLE of $\boldsymbol{\theta}_0$ is the maximizer $\hat{\boldsymbol{\theta}}_n$ of $l_n(\boldsymbol{\theta}_n; \text{data})$ over Ψ_n given by (2.9).

2.2 A Nonparametric Association Test

Suppose that $F_{0,0}(t_1, t_2)$ is the underlying joint distribution function of T_1 and T_2, and $F_{0,1}(t_1)$ and $F_{0,2}(t_2)$ are the underlying marginal distribution functions for T_1 and T_2, respectively. It is noted that

$$F_{0,0}(t_1, t_2) = F_{0,1}(t_1) F_{0,2}(t_2) \quad \text{for any} \quad (t_1, t_2) \in [\tau_{1,l}, \tau_{1,h}] \times [\tau_{2,l}, \tau_{2,u}],$$

if T_1 and T_2 are independent. It naturally leads to consider a functional of the distribution functions $\boldsymbol{\theta}_0 = (F_{0,0}(\cdot,\cdot), F_{0,1}(\cdot), F_{0,2}(\cdot))$,

$$\rho(\boldsymbol{\theta}_0) = \int_{\tau_{1,l}}^{\tau_{1,h}} \int_{\tau_{2,l}}^{\tau_{2,h}} \{F_{0,0}(t_1, t_2) - F_{0,1}(t_1) F_{0,2}(t_2)\}\, dt_2 dt_1$$

as the basis to construct the statistic for testing the association between T_1 and T_2. We propose to study the test statistic,

$$\rho(\hat{\boldsymbol{\theta}}_n) = \int_{\tau_{1,l}}^{\tau_{1,h}} \int_{\tau_{2,l}}^{\tau_{2,h}} \left\{ \hat{F}_{n,0}(t_1, t_2) - \hat{F}_{n,1}(t_1)\hat{F}_{n,2}(t_2) \right\} dt_2 dt_1, \tag{2.10}$$

computed in a two-stage approach, where $\hat{\boldsymbol{\theta}}_n$ is the spline-based sieve NPMLE described above.

Under H_0: T_1 and T_2 are independent, $\rho(\boldsymbol{\theta}_0) = 0$ and hence it is anticipated that $\rho(\hat{\boldsymbol{\theta}}_n)$ is asymptotically close to zero. Moreover, the forthcoming Theorem 2 will justify $\sqrt{n}(\rho(\hat{\boldsymbol{\theta}}_n) - \rho(\boldsymbol{\theta}_0))$ converges in distribution to a normal variable with mean 0. This leads to the construction of the standard normal test statistic $T_n = \rho(\hat{\boldsymbol{\theta}}_n)/SE(\rho(\hat{\boldsymbol{\theta}}_n))$ for H_0, where $SE(\rho(\hat{\boldsymbol{\theta}}_n))$ can be estimated by the bootstrap method.

3. Asymptotic Theorems

In this section, we develop consistency and the rate of convergence theorem for the proposed sieve NPMLE. Furthermore, we establish the asymptotic normality theorem for the proposed nonparametric functional test statistics for the association and demonstrate its efficiency. Study of the asymptotic properties needs empirical process theory and requires some regularity conditions regarding the event and observation times. For the theoretical development throughout this paper, let c be a positive constant that may have

different values from place to place. The following conditions sufficiently guarantee the results in the forthcoming theorem for the consistency and rate of convergence for the proposed sieve NPMLEs, and will be used to establish the asymptotic normality for the proposed association test statistic and to demonstrate its efficiency. For the sake of easy in notation, let $D^{\alpha} = \frac{\partial^{[\alpha]}}{\partial t_1^{\alpha_1} \partial t_2^{\alpha_2}}$ with $[\alpha] = \alpha_1 + \alpha_2$ for nonnegative integers α_1 and α_2.

Regularity Conditions:

C1 For every α with $[\alpha] < p$, $D^{\alpha} F_{0,0}(t_1, t_2)$ is continuous at any (t_1, t_2) in $[0, \tau_1] \times [0, \tau_2]$. Moreover for $[\alpha] = p$, $D^{\alpha} F_{0,0}(t_1, t_2)$ exists and satisfies $|D^{\alpha} F_{0,0}(t_1, t_2) - D^{\alpha} F_{0,0}(t_1', t_2')| \leq c(|t_1 - t_1'|^r + |t_2 - t_2'|^r)$ for $r > 0$.

C2 $F_{0,1}(t_1)$ and $F_{0,2}(t_2)$ both have up to $(p-1)$th continuous derivatives on $[0, \tau_1]$ and $[0, \tau_2]$, respectively. In addition, their pth derivatives also exist and satisfy $|d^p F_{0,1}(t_1)/dt_1^p - d^p F_{0,1}(t_1')/dt_1^p| \leq c|t_1 - t_1'|^r$ and $|d^p F_{0,2}(t_2)/dt_2^p - d^p F_{0,1}(t_2')/dt_2^p| \leq c|t_2 - t_2'|^r$, where p and r are the same as in C1.

C3 $\frac{\partial^2 F_{0,0}(t_1,t_2)}{\partial t_1 \partial t_2}$ have a positive lower bound in $[0, \tau_1] \times [0, \tau_2]$.

C4 The joint density of (U_1, V_1, U_2, V_2) is continuous and has a positive lower bound in its domain $\mathcal{D}$ with $\mathcal{D}$ defined by (2.7).

Let $\boldsymbol{\theta}_1 = (F_{1,0}(\cdot,\cdot), F_{1,1}(\cdot), F_{1,2}(\cdot))$ and $\boldsymbol{\theta}_2 = (F_{2,0}(\cdot,\cdot), F_{2,1}(\cdot), F_{2,2}(\cdot))$. Define $d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ as

$$\begin{aligned} d^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) =& \|F_{1,0} - F_{2,0}\|^2_{L_2(P_{U_1,U_2})} + \|F_{1,0} - F_{2,0}\|^2_{L_2(P_{U_1,V_2})} \\ &+ \|F_{1,0} - F_{2,0}\|^2_{L_2(P_{V_1,U_2})} + \|F_{1,0} - F_{2,0}\|^2_{L_2(P_{V_1,V_2})} \\ &+ \|F_{1,1} - F_{2,1}\|^2_{L_2(P_{U_1})} + \|F_{1,1} - F_{2,1}\|^2_{L_2(P_{V_1})} \\ &+ \|F_{1,2} - F_{2,2}\|^2_{L_2(P_{U_2})} + \|F_{1,2} - F_{2,2}\|^2_{L_2(P_{V_2})}, \end{aligned} \tag{3.11}$$

where each of the L_2-norms is associated with a specific probability measure. For example, $\|\cdot\|_{L_2(P_{U_1,U_2})}$ is the L_2-norm associated with the true probability measure P_{U_1,U_2} of (U_1, U_2).

Theorem 1. *Suppose that C1–C4 hold, $p_n = O(n^\kappa)$ and $q_n = O(n^\kappa)$ for p_n and q_n used in (2.2) and (2.3). Then there exists a sub set $\Theta_n \subset \Psi_n$ for Ψ_n defined by (2.9), such that for $l_n(\cdot; data)$ defined by (2.1), the maximizer $\hat{\boldsymbol{\theta}}_n$ of $l_n(\boldsymbol{\theta}_n; data)$ over Θ_n is a consistent estimator of the vector of underlying distribution functions $\boldsymbol{\theta}_0 = (F_{0,0}, F_{0,1}, F_{0,2})$ and*

$$d\left(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0\right) = O_P\left(n^{-\min\{(p+r)\kappa, (1-2\kappa)/2\}}\right).$$

Theorem 1 implies that the optimal rate of convergence for $\hat{\boldsymbol{\theta}}_n$ is $n^{\frac{p+r}{2(p+r+1)}}$, achieved when κ is chosen as $\frac{1}{2(p+r+1)}$. This rate is known as the optimal rate for bivariate nonparametric regression problem (Stone, 1982), though slower than $n^{1/2}$ even for large $p + r$.

Although the proposed sieve NPMLE does not achieve the convergence rate of $n^{1/2}$, it can be shown that the proposed test statistic $\rho(\hat{\boldsymbol{\theta}}_n)$ as a functional of the NPMLEs can still be asymptotically normal with ordinary convergence rate of $n^{1/2}$ by Riesz representation theorem for Hilbert space (Halmos, 1982). Note that similar ideas were adopted by Shen (1997) and Chen et al. (2006) for relatively simple estimation problems, but our proof for the normality (Theorem 2) is more technically challenged because of the complicated nature in estimating the joint distribution function.

Define

$$\mathfrak{W} = \{\boldsymbol{w} = (w_0(\cdot,\cdot), w_1(\cdot), w_2(\cdot)) : \boldsymbol{w} \text{ being a vector of piecewise continuous functions with bounded derivatives for } \frac{\partial^2 w_0(t_1,t_2)}{\partial t_1 \partial t_2}, \frac{dw_1(t_1)}{dt_1} \text{ and } \frac{dw_2(t_2)}{dt_2}\Big\}.$$

Since continuity implies piecewise continuity, the vector of the target distribution function $\boldsymbol{\theta}_0$ belongs to $\mathfrak{W}$. Let $\boldsymbol{X} = \left\{U_1, V_1, U_2, V_2, \left(\Delta_1^{(j)}, \Delta_2^{(j)}\right)_{j=1}^{3}\right\}$, a random observation for bivariate interval-censored event times. For model parameter $\boldsymbol{\theta} = \{F_0(\cdot,\cdot), F_1(\cdot), F_2(\cdot)\}$, the log likelihood function of $\boldsymbol{\theta}$ given

a single observation $\boldsymbol{X}$ is

$$\begin{aligned} l(\boldsymbol{\theta};\boldsymbol{X}) =& \Delta_1^{(1)}\Delta_2^{(1)}\log F_0(U_1,U_2) + \Delta_1^{(1)}\Delta_2^{(2)}\log\{F_0(U_1,V_2)-F_0(U_1,U_2)\} \\ &+ \Delta_1^{(1)}\Delta_2^{(3)}\log\{F_1(U_1)-F_0(U_1,V_2)\} \\ &+ \Delta_1^{(2)}\Delta_2^{(1)}\log\{F_0(V_1,U_2)-F_0(U_1,U_2)\} \\ &+ \Delta_1^{(2)}\Delta_2^{(2)}\log\{F_0(V_1,V_2)-F_0(V_1,U_2)-F_0(U_1,V_2)+F_0(U_1,U_2)\} \\ &+ \Delta_1^{(2)}\Delta_2^{(3)}\log\{F_1(V_1)-F_0(V_1,V_2)-F_1(U_1)+F_0(U_1,V_2)\} \\ &+ \Delta_1^{(3)}\Delta_2^{(1)}\log\{F_2(U_2)-F_0(V_1,U_2)\} \\ &+ \Delta_1^{(3)}\Delta_2^{(2)}\log\{F_2(V_2)-F_2(U_2)-F_0(V_1,V_2)+F_0(V_1,U_2)\} \\ &+ \Delta_1^{(3)}\Delta_2^{(3)}\log\{1-F_2(V_2)-F_1(V_1)+F_0(V_1,V_2)\}. \end{aligned}$$

Then for $\boldsymbol{w},\tilde{\boldsymbol{w}} \in \mathfrak{W}$, the first directional derivative along $\boldsymbol{w}$ and the second directional derivative along $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ of $l(\boldsymbol{\theta};\boldsymbol{X})$ evaluated at $\boldsymbol{\theta}_0$ are, respectively, given by

$$\frac{dl(\boldsymbol{\theta}_0;\boldsymbol{X})}{d\boldsymbol{\theta}}[\boldsymbol{w}] \equiv \frac{dl(\boldsymbol{\theta}_0+s\boldsymbol{w};\boldsymbol{X})}{ds}\mid_{s=0}, \tag{3.12}$$

and

$$\frac{d^2l(\boldsymbol{\theta}_0;\boldsymbol{X})}{d\boldsymbol{\theta}^2}[\boldsymbol{w}][\tilde{\boldsymbol{w}}] \equiv \frac{d\left\{\frac{dl(\boldsymbol{\theta}_0+s\tilde{\boldsymbol{w}};\boldsymbol{X})}{d\boldsymbol{\theta}}[\boldsymbol{w}]\right\}}{ds}\mid_{s=0}. \tag{3.13}$$

Note that by the regularity conditions C1–C3 and the construction of $\mathfrak{W}$, for any $\boldsymbol{w} \in \mathfrak{W}$, there exists a small neighbourhood of 0, such that for each s in this neighbourhood, $\boldsymbol{\theta}_0 + s\boldsymbol{w}$ is also a vector of distribution

functions formed by a joint distribution function and its two corresponding marginal distribution functions, and that $l(\boldsymbol{\theta}_0 + s\boldsymbol{w}; \boldsymbol{X})$ is bounded. Hence, the directional derivatives (3.12) and (3.13) are both well defined.

Let P be the probability measure of $\boldsymbol{X}$ with the underlying model parameter $\boldsymbol{\theta}_0$. Based on the directional derivative, the Fisher information inner product for $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ is defined as

$$\langle \boldsymbol{w}, \tilde{\boldsymbol{w}} \rangle = P\left\{ \left(\frac{dl(\boldsymbol{\theta}_0; \boldsymbol{X})}{d\boldsymbol{\theta}}[\boldsymbol{w}] \right) \left(\frac{dl(\boldsymbol{\theta}_0; \boldsymbol{X})}{d\boldsymbol{\theta}}[\tilde{\boldsymbol{w}}] \right) \right\}$$

and the Fisher information norm for $\boldsymbol{w}$ is given by

$$\|\boldsymbol{w}\|^2 = \langle \boldsymbol{w}, \boldsymbol{w} \rangle. \tag{3.14}$$

For any $\boldsymbol{w} \in \mathfrak{W}$, we write

$$\frac{d\rho(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}[\boldsymbol{w}] \equiv \frac{d\rho(\boldsymbol{\theta}_0 + s\boldsymbol{w})}{ds} \mid_{s=0} . \tag{3.15}$$

Then, it immediately follows that

$$\begin{aligned} \frac{d\rho(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}[\boldsymbol{w}] &= \lim_{s \to 0} \frac{\rho(\boldsymbol{\theta}_0 + s\boldsymbol{w}) - \rho(\boldsymbol{\theta}_0)}{s} \\ &= \int_{\tau_{1,l}}^{\tau_{1,h}} \int_{\tau_{2,l}}^{\tau_{2,h}} \{ w_0(t_1, t_2) - F_{0,1}(t_1) w_2(t_2) - w_1(t_1) F_{0,2}(t_2) \} \, dt_2 dt_1. \end{aligned} \tag{3.16}$$

Theorem 2. *Given that C1–C4 hold and $p + r > 3$ in C1 and C2,*

$$\sqrt{n} \left\{ \rho(\hat{\boldsymbol{\theta}}_n) - \rho(\boldsymbol{\theta}_0) \right\} \to_d N\left(0, \left\| \frac{d\rho(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}} \right\|_{*,\infty}^2 \right),$$

where $\left\| \frac{d\rho(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}} \right\|_{*,\infty} = \sup_{\boldsymbol{w} \in \mathfrak{W}, \|\boldsymbol{w}\| > 0} \frac{\left| \frac{d\rho(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}[\boldsymbol{w}] \right|}{\|\boldsymbol{w}\|}$.

Based on Theorem 2, we know $\frac{\sqrt{n}\{\rho(\hat{\boldsymbol{\theta}}_n)-\rho(\boldsymbol{\theta}_0)\}}{\left\|\frac{d\rho(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}\right\|_{*,\infty}}$ converges in distribution to the standard normal distribution. But in view of (3.16), the direct estimation of $\left\|\frac{d\rho(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}\right\|_{*,\infty}$ is not straightforward. So the nonparametric bootstrap method is recommended to estimate the standard error for the test statistic described in Section 2.2.

Next we investigate the efficiency of the proposed estimator $\rho\left(\hat{\boldsymbol{\theta}}_n\right)$. First, we define the path-wise regular estimator for $\rho\left(\boldsymbol{\theta}_0\right)$ according to the concept originally proposed by Wong (1992). An estimator T_n for $\rho(\boldsymbol{\theta}_0)$ is called path-wise regular, if $\sqrt{n}\{T_n - \rho(\boldsymbol{\theta}_0)\}$ converges in distribution to a normal distribution; and if for any $h > 0$ and any $\boldsymbol{w} \in \mathfrak{W}$ with $\|\boldsymbol{w}\| > 0$ and $s_n \to 1$ we have

$$\limsup \Pr_{\boldsymbol{\theta}_{n,h}}\left\{T_n < \rho\left(\boldsymbol{\theta}_{n,h}\right)\right\} \leq \liminf \Pr_{\boldsymbol{\theta}_{n,-h}}\left\{T_n < \rho\left(\boldsymbol{\theta}_{n,-h}\right)\right\},$$

where $\boldsymbol{\theta}_{n,h} = \boldsymbol{\theta}_0 + \frac{s_n h}{\sqrt{n}}\boldsymbol{w}$ and $\boldsymbol{\theta}_{n,-h} = \boldsymbol{\theta}_0 - \frac{s_n h}{\sqrt{n}}\boldsymbol{w}$, $\Pr_{\boldsymbol{\theta}_{n,h}}$ and $\Pr_{\boldsymbol{\theta}_{n,-h}}$ denote the probabilities, respectively, under the probability measures $P_{\boldsymbol{\theta}_{n,h}}$ and $P_{\boldsymbol{\theta}_{n,-h}}$.

Note that for each $h > 0$, both $\boldsymbol{\theta}_{n,h}$ and $\boldsymbol{\theta}_{n,-h}$ are well defined vector of distribution functions when n is large enough, so both probability measures are also well defined for large sample.

Theorem 3. *Given that C1–C4 hold and $p + r > 3$ in C1 and C2, the*

proposed plug-in estimator $\rho\left(\hat{\boldsymbol{\theta}}_n\right)$ *is the optimal path-wise regular estimator for* $\rho(\boldsymbol{\theta}_0)$. *That is, the asymptotic variance for* $\rho\left(\hat{\boldsymbol{\theta}}_n\right)$ *reaches the lower bound for all path-wise regular estimators for* $\rho(\boldsymbol{\theta}_0)$.

Theorem 3 implies that our proposed association test is the most powerful test based on $\rho\left(\hat{\boldsymbol{\theta}}_n\right)$ among all path-wise regular estimators for $\rho(\boldsymbol{\theta}_0)$.

4. Computation of the Sieve NPMLE

The proposed sieve NPMLE is restricted to Ψ_n defined in (2.9), Section 2.1. For a given set of spline knots, it leads to determine the coefficients of the spline estimate with the resulting spline-based sieve estimate $\hat{\boldsymbol{\theta}}_n$ maximizing the log likelihood of (2.1) over Ψ_n. It is, however, still a numerically daunting task. We note that restricting $\boldsymbol{\theta}_n$ inside Ψ_n, it is sufficient that

the spline coefficients for (2.4), (2.5) and (2.6) satisfy:

$$\begin{aligned}
&\alpha_{i,1} = 0 \text{ for } i = 1, \ldots, p_n, \\
&\alpha_{1,j} = 0 \text{ for } j = 2, \ldots, q_n, \\
&(\alpha_{i+1,j+1} - \alpha_{i+1,j}) - (\alpha_{i,j+1} - \alpha_{i,j}) \geq 0 \\
&\qquad \text{for } i = 1, \ldots, p_n - 1, j = 1, \ldots, q_n - 1, \\
&\beta_1 = 0, \ \gamma_1 = 0, \\
&(\beta_{i+1} - \beta_i) - (\alpha_{i+1,q_n} - \alpha_{i,q_n}) \geq 0 \text{ for } i = 1, \ldots, p_n - 1, \\
&(\gamma_{j+1} - \gamma_j) - (\alpha_{p_n,j+1} - \alpha_{p_n,j}) \geq 0 \text{ for } j = 1, \ldots, q_n - 1, \\
&\beta_{p_n} + \gamma_{q_n} - \alpha_{p_n,q_n} \leq 1.
\end{aligned} \tag{4.17}$$

Therefore, we suggest to compute the spline-based sieve NPMLE inside a subset of Ψ_n, Ψ_n' given by

$$\Psi_n' = \Big\{ \boldsymbol{\theta}_n = (F_{n,0}, F_{n,1}, F_{n,2}) : F_{n,0}(\cdot,\cdot) = \sum_{i=1}^{p_n}\sum_{j=1}^{q_n} \alpha_{i,j} B_i^{(1),l}(\cdot) B_j^{(2),l}(\cdot),$$
$$F_{n,1}(\cdot) = \sum_{i=1}^{p_n} \beta_i B_i^{(1),l}(\cdot), F_{n,2}(\cdot) = \sum_{j=1}^{q_n} \gamma_j B_j^{(2),l}(\cdot),$$
$$\text{(4.17) holds, knot sequences are as (2.2) and (2.3)} \Big\}$$

Since (4.17) is quite complicated for numerical implementation, we used I-splines instead of B-splines for computation following the same approach as in Wu and Zhang (2012). Let I_i^l denote the I-spline of degree l associated with the ith knot defined by Ramsay (1988), it is known that

$I_i^l(t) = \sum_{h=i+1}^{p_n+1} B_h^{l+1}(t)$. Then some straightforward algebra yields that $F_n(\cdot,\cdot)$, $F_{n,1}(\cdot)$ and $F_{n,2}(\cdot)$ given by (2.4), (2.5) and (2.6) with constraints (4.17) are equivalent to

$$F_n(\cdot,\cdot) = \sum_{i=1}^{p_n-1}\sum_{j=1}^{q_n-1} \mu_{i,j} I_i^{(1),l-1}(\cdot) I_j^{(2),l-1}(\cdot), \tag{4.18}$$

$$F_{n,1}(\cdot) = \sum_{i=1}^{p_n-1} \left\{ \sum_{j=1}^{q_n-1} \mu_{i,j} + \omega_i \right\} I_i^{(1),l-1}(\cdot) \tag{4.19}$$

and

$$F_{n,2}(\cdot) = \sum_{j=1}^{q_n-1} \left\{ \sum_{i=1}^{p_n-1} \mu_{i,j} + \pi_j \right\} I_j^{(2),l-1}(\cdot) \tag{4.20}$$

with constraints

$$\begin{aligned} &\mu_{i,j} \geq 0 \text{ for } i = 1, \cdots, p_n - 1, j = 1, \cdots, q_n - 1, \\ &\omega_i \geq 0, i = 1, \cdots, p_n - 1, \\ &\pi_j \geq 0, j = 1, \cdots, q_n - 1, \\ &\sum_{i=1}^{p_n-1}\sum_{j=1}^{q_n-1} \mu_{i,j} + \sum_{i=1}^{p_n-1} \omega_i + \sum_{j=1}^{q_n-1} \pi_j \leq 1. \end{aligned} \tag{4.21}$$

That is Ψ_n' can be written as

$$\Psi_n' = \Bigg\{ \boldsymbol{\theta}_n = (F_{n,0}, F_{n,1}, F_{n,2}) : F_{n,0}(\cdot,\cdot) = \sum_{i=1}^{p_n-1}\sum_{j=1}^{q_n-1} \mu_{i,j} I_i^{(1),l-1}(\cdot) I_j^{(2),l-1}(\cdot),$$
$$F_{n,1}(\cdot) = \sum_{i=1}^{p_n-1} \left\{ \sum_{j=1}^{q_n-1} \mu_{i,j} + \omega_i \right\} I_i^{(1),l-1}(\cdot),$$
$$F_{n,2}(\cdot) = \sum_{j=1}^{q_n-1} \left\{ \sum_{i=1}^{p_n-1} \mu_{i,j} + \pi_j \right\} I_j^{(2),l-1}(\cdot),$$
$$\text{(4.21) holds, knot sequences are as (2.2) and (2.3)} \Bigg\}. \tag{4.22}$$

Let $\boldsymbol{\mu} = \{\mu_{i,j}\}_{i=1,\cdots,p_n-1, j=1,\cdots,q_n-1}$, $\boldsymbol{\omega} = \{\omega_i\}_{i=1,\cdots,p_n-1}$ and $\boldsymbol{\pi} = \{\pi_j\}_{j=1,\cdots,q_n-1}$. The log likelihood (2.1) in the sieved space Ψ_n', $l_n(\boldsymbol{\theta}_n; \text{data})$ can be written as $l(\boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{\pi}; \text{data})$ and treated as a function of $(\boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{\pi})$. Then the sieve NPMLE can be obtained through finding the maximizer $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\pi}})$ of $l(\boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{\pi}; \text{data})$ subject to the constraints (4.21). The generalized gradient projection method (Jamshidian, 2004) can be used for this constrained maximization problem. The detailed algorithm can also be found in Wu and Zhang (2012) and Zhang et al. (2010).

The knot selection is an important step for implementing the spline-based sieve estimation. In this work, the spline knot sequence in T_1 direction was chosen based on the quantile of observation times $\mathcal{O} = \{(u_{1,k}, v_{1,k})\}_{k=1}^n$. We selected the number of the interior knots to be $[n^{1/3}]$ (the closest integer to $n^{1/3}$), and put interior knots at quantiles of $\mathcal{O}$. The same knot sequence

selection procedure was applied to T_2 direction.

5. Simulation studies

We conducted simulations studies for bivariate event time data generated from Clayton copula model (Clayton, 1978):

$$F_{0,0}(t_1, t_2) = \left\{F_{0,1}^{-\alpha}(t_1) + F_{0,2}^{-\alpha}(t_2) - 1\right\}^{-\frac{1}{\alpha}},$$

in which α indicates the association between T_1 and T_2 as the association measure $Kendall's\ \tau$ for Clayton copula is related to α by $Kendall's\ \tau = \alpha/(\alpha+2)$. For simulation studies, we let marginal distribution of both event times follow the exponential distribution with hazard 0.5, which results in the cumulative distribution function $F_{0,i}(t) = 1 - \exp(-0.5t)$ $(i = 1, 2)$. We consider four scenarios of correlated bivariate event times with $\alpha = 0.222, 0.667, 2,$ and 6 that correspond to $Kendall's\ \tau = 0.1, 0.25, 0.5,$ and 0.75, respectively, in addition to the scenario of uncorrelated bivariate event times. For interval censoring, we let U_i and V_i for $i = 1, 2$ all follow uniform distribution over interval $[0.0201, 4.7698]$, which yields a small probability of the event time falling outside this range ($\Pr(0 < T_i < 0.0201) = \Pr(4.7698 < T_i < 5) = 0.01$ for $i = 1, 2.$) and we also restricted the censoring interval satisfying $V_i - U_i > 0.05$ $(i = 1, 2)$.

We generated interval-censored bivariate event time data for each of

the five scenarios described above with sample size of 100 and 200, respectively. For each data sample, we adopted cubic I-splines ($l = 3$) to compute the sieve NPMLEs with the interior knots selected in the way outlined in Section 4 and the boundary knots were chosen as 0 and $\max\{\mathcal{O}\} + 0.5$, 0.5 to the right of the largest observation time. For each scenario, we repeated the experiment 1,000 times to evaluate the estimation performance. For all these simulation scenarios, the percentages of left-, interval, and right-censored observations are roughly 48%, 28% and 24%, respectively for both event times T_1 and T_2.

We plotted the bias of the sieve NPMLE of the joint distribution function $F_{0,0}$ for the two cases of uncorrelated bivariate event times and correlated bivariate event times with $Kendall's\ \tau = 0.75$ with sample size 200 in Figure 1. It appears from the plots that the sieve NPMLE of the joint distribution based on 200 observations had virtually ignorable estimation bias having maximum point-wise bias of 0.0074 and 0.0412, respectively, for uncorrelated bivariate event times case and correlated bivariate event times case with $Kendall's\ \tau = 0.75$. The results for other three scenarios (not shown here) were similar with the maximum point-wise estimation bias between the two values.

For the association test, we computed the plug-in estimate

$$\rho(\hat{\boldsymbol{\theta}}_n) = \int_{0.1}^{4.0} \int_{0.1}^{4.0} \left\{ \hat{F}_{n,0}(t_1, t_2) - \hat{F}_{n,1}(t_1)\hat{F}_{n,2}(t_2) \right\} dt_2 dt_1,$$

the efficient path-wise regular estimator of

$$\rho(\boldsymbol{\theta}_0) = \int_{0.1}^{4.0} \int_{0.1}^{4.0} \{F_{0,0}(t_1, t_2) - F_{0,1}(t_1)F_{0,2}(t_2)\} dt_2 dt_1,$$

by Theorem 3. Table 1 presents the simulation results for the association test based on the plug-in estimate of $\rho(\boldsymbol{\theta}_0)$, including the mean of $\rho\left(\hat{\boldsymbol{\theta}}_n\right)$, Monte-Carlo standard deviation (MCSD), mean of the estimated standard errors (BSE) based on 100 bootstrap samples, the 95% coverage probability (CP) with 95% Wald confidence interval, and the rejection probability (RP) for testing the null hypothesis H_0: two event times are independent, at a significance level of 0.05 based on $1,000$ repetitions. The results show that the finite sample performance of the proposed test statistic based on asymptotic normality theorem established in Theorem 2 is quite satisfactory. The estimation bias is ignorable, the BSE is valid because the mean of BSEs is quite close to the MCSD, the CP is around its nominal value of 0.95 for all scenarios even with sample size of 100. Moreover, the proposed association test has the right size of 0.05 for independent bivariate event time case and a decent power to detect the association between correlated bivariate event times. For bivariate event times with $Kendall's\ \tau = 0.25$,

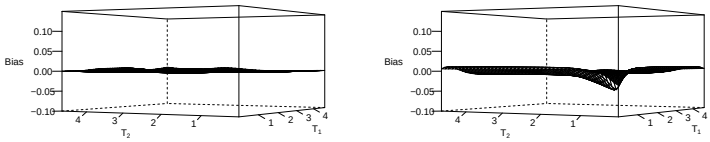

Figure 1: The average sieve estimation bias for $F_{(0,0)}$ for sample size 200 with uncorrelated (left panel) and $Kendall's\ \tau = 0.75$ (right panel)

the test has 86% power to reject the H_0 with sample size of 100 and it has almost 100% power to reject the H_0 when $Kendall's\ \tau$ is 0.5 or larger under this simulation setting even with sample size of 100. We found that CP deceases a little as Kendall's τ increases. This is likely due to the fact that large Kendall's τ tends to result in relatively large finite sample estimation bias for the joint distribution. With that being said, the empirical power seemingly increases as τ increases from the simulation studies for the objective of association test. These results for empirical power are expected since smaller Kendall's τ means weaker association and smaller effect size, which gives lower power for our association test.

It is worth noting that the computing time for this spline-based non-parametric analysis is very manageable with the major effort spent on the computation of sieve NPMLE. Though the numerical algorithm appears to be very complicated because of the constraints, it only took on average

Table 1: The results for the simulation studies on the association test for all scenarios with sample size 100 and 200

Sample Size	$\rho(\boldsymbol{\theta}_0)$	$\rho(\hat{\boldsymbol{\theta}}_n)$	MCSD	BSE	CP	RP
			Scenario 1: Uncorrelated			
100	0	0.016	0.239	0.242	0.944	0.056
200	0	0.005	0.168	0.171	0.955	0.045
			Scenario 2: $Kendall's\ \tau = 0.10$			
100	0.209	0.219	0.250	0.240	0.934	0.158
200	0.209	0.203	0.173	0.170	0.950	0.228
			Scenario 3: $Kendall's\ \tau = 0.25$			
100	0.525	0.508	0.243	0.235	0.944	0.570
200	0.525	0.511	0.165	0.166	0.944	0.860
			Scenario 4: $Kendall's\ \tau = 0.50$			
100	1.042	1.015	0.233	0.227	0.934	0.994
200	1.042	1.006	0.158	0.157	0.937	1.000
			Scenario 5: $Kendall's\ \tau = 0.75$			
100	1.506	1.424	0.208	0.208	0.928	1.000
200	1.506	1.460	0.144	0.147	0.930	1.000

about 5.3 seconds to complete the computation for data with sample size 200 using a Lenovo ThinkPad with Intel Core I5-5300U CPU. The computing algorithm was implemented in R and will be available from the first author upon request. In addition, the test results appear not be impacted much by the selection of integral limits. We have tried other integral limits for the defnition of $\rho(\theta)$, while the estimate of $\rho(\hat{\theta}_n)$ depends on the choice of limits, the Wald-test statistic seems insensitive to the selection of the limits that results in very similar rejection probability.

6. Real data analysis

HD is an autosomal dominant neurodegenerative disease caused by expansion of the trinucleotide cytosine-adenine-guanine (CAG) in the huntington gene (Walker, 2007). The Neurobiological Predictors of Huntington's Disease (PREDICT-HD) project was a 12-year prospective observational cohort study from 2002 to 2014 on premanifest-HD individuals for HD progression with a goal to identify useful clinical and biological markers that are predictive of the landmark event, clinical motor diagnosis of HD (Paulsen et al., 2014). Cognitive impairment as one of the "triad" of clinical symptoms (motor, cognitive, psychiatric) has been often the study of interest in HD. Mild cognitive impairment (MCI), as a clinically diagnostic entity,

has been recognized as a translational phase between normal aging and dementia and becomes increasingly significant as a study endpoint to clinical trials in treating neurodegenerative diseases, see for example Petersen (2004), Caviness et al. (2007) and Duff et al. (2010). For HD, Harrington et al. (2012) identified six cognitive domains consisting of Speed & Inhibition, Verbal Working Memory, Motor Planning, Attention & Information Integration, Sensory & Perceptual, and Verbal Learning & Memory. The cognitive impairment in Verbal Working Memory or Verbal Learning & Memory is regarded as amnestic and the impairment on other domains is non-amnestic (Duff et al., 2010). We denote the MCI diagnosed in amnestic and non-amnestic areas as MCI-A and MCI-NA, respectively. We applied the proposed method to test the association between the ages of onset for MCI-A and MCI-NA in premanifest-HD individuals using data from the PREDICT-HD study.

There were a total of 799 premanifest-HD individuals available for ascertaining cognitive impairment in both amnestic and non-amnestic domains through periodic assessments in a battery of neuropsychological tests in the PREDICT-HD study. It provided bivariate interval-censored observation of the ages of onset for both MCI subtypes. Table 2 summarizes the bivariate interval-censored MCI events in both subtypes. It appears that the

non-amnestic MCI was more frequently diagnosed than the amnestic MCI during the study period, which is consistent with the observation by Duff et al. (2010).

Table 2: Summary of interval-censored MCI events in the PREDICT-HD study

Cognitive Domains	Left-Censored	Interval-Censored	Right-Censored	Total
Amnestic	201	98	500	799
Non-Amnestic	365	208	226	799

Figure 2 plots the proposed spline-based sieve NPMLEs of the distribution functions of the ages of onset and reports the 1st quantile, median, and the 3rd quantile for the ages of onset for both MCI subtypes. As expected, the non-amnestic MCI can occur much earlier than the amnestic MCI with the estimated median onset ages being 39.2 and 58.8, respectively, which explains why MCI-NA was more frequently diagnosed than MCI-A in the PREDICT-HD study.

We applied the proposed association test to examine the association between the two MCI subtypes using the estimated functional

$$\rho(\hat{\boldsymbol{\theta}}_n) = \int_{20}^{80} \int_{20}^{80} \left\{ \hat{F}_{n,0}(t_1, t_2) - \hat{F}_{n,1}(t_1)\hat{F}_{n,2}(t_2) \right\} dt_2 dt_1.$$

The test statistic $T_n = \rho(\hat{\boldsymbol{\theta}}_n)/\text{BSE}$ with BSE given by 100 bootstrap re-

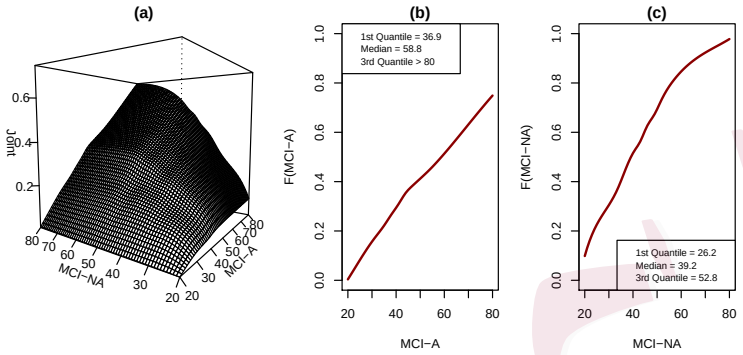

Figure 2: The spline-based sieve NPMLEs of joint and marginal distribution functions of the ages of onset for both MCI subtypes

samples was 9.26 yielding $p - value < 0.00001$. It implies that both MCI subtypes are strongly correlated. For the estimated functional, the integral limits were selected according to the age span in the PREDICT-HD study to make sure that there are some portion of censoring time points below the lower integral limit as well as above the upper integral limit. In general, a small portion between 1% to 10% will suffice. As observed in the simulation studies, the test results are quite robust against the choice of integral limits. The test statistic $T_n's$ were all around 9 when the limits were chosen to be (30,70) and (40,60).

7. Concluding remarks

The analysis of bivariate interval-censored data is a very challenging problem both computationally and theoretically. To the best of our knowledge, the proposed spline-based nonparametric method is the first complete and theoretically justified approach without any distributional structure assumed for bivariate event times in the published literature. Not only the spline-based sieve NPMLEs for both joint and marginal distribution functions enjoy the estimation consistency and the optimal rate of convergence in bivariate nonparametric regression, the proposed model-free association test is also shown to be a powerful test with the ordinary asymptotic normality property. Our simulation studies demonstrated the superior finite sample performance of the proposed method as well as its numerical advantage for such a challenging problem.

Though the proposed association test works very well for the simulation settings in our numerical experiments, it is worth investigating if a weighted test based on the functional

$$\rho_w(\theta) = \int_{\tau_{1,l}}^{\tau_{1,h}} \int_{\tau_{2,l}}^{\tau_{2,h}} w(t_1, t_2)\{F_0(t_1, t_2) - F_1(t_1)F_2(t_2)\}dt_2 dt_1$$

could improve the power of the test with the optimal choice of the weight function $w(t_1, t_2)$ for a given situation. While testing the association be-

tween the two event times is important in many applications, it is also desired to be able to evaluate the global association quantitatively. Having the superior estimation properties in the spline-based sieved NPMLEs and the asymptotic normality of the functionals of $\hat{\theta}_n = (\hat{F}_{n,0}(\cdot,\cdot), \hat{F}_{n,1}(\cdot), \hat{F}_{n,2}(\cdot))$, one may consider to explore direct estimation of correlation of coefficient

$$\tau(T_1, T_2) = \frac{Cov(T_1, T_2)}{\sqrt{Var(T_1)Var(T_2)}},$$

in which both numerator and denominator can be expressed as some smooth functionals of $\theta = (F_0(\cdot,\cdot), F_1(\cdot), F_2(\cdot))$. However, the asymptotic properties for such a plug-in functional estimator are not easy to study and leave us as an interesting inference problem for future research. Another area for assessing the association between bivariate event times is to study the time-dependent cross ratio. Nan et al. (2006) and Hu et al. (2011) developed some nonparametric procedures to estimate the time-independent cross ratio function with bivariate right-censored event time data.

Supplementary Material

The supplementary material contains technical details including lemmas and their proofs necessary for the main paper.

Acknowledgements

The research of Yuan Wu was supported in part by award number

P01CA142538 from the National Cancer Institute. The research of Ying Zhang and Junyi Zhou was supported in part by award number R01NS103475 from National Institute of Neurological Disorders and Stroke. The content is solely the responsibility of the authors and does not necessarily represent the official views of National Institute of Health.

References

Betensky, R. A. and D. M. Finkelstein (1999). A nonparametric maximum likelihood estimator for bivariate censored data. *Statistics in Medicine 18*, 3089–3100.

Bollaerts, K., P. H. C. Eilers, and I. van Mechelen (2006). Simple and multiple p-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology 59*, 451–469.

Brown, B. W. M., M. Hollander, and R. M. Korwar (1974). Nonparametric tests of independence for censored data with applications to heart transplant studies. *Reliability and Biometry*, 327–354.

Caviness, J., E. Driver-Dunckley, D. Connor, M. Sabbagh, J. Hentz, B. Noble, Evidente, H. VG. Shill, and C. Adler (2007). Defining mild cognitive impairment in pakinson's disease. *Movement Disorders 22*, 1272–1277.

Chen, X., Y. Fan, and V. Tsyrennikov (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association 475*, 1228–1240.

REFERENCES

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*, 141–151.

Ding, A. A. and W. Wang (2004). Testing independence for bivariate current status data. *Journal of the American Statistical Association 99*, 145–155.

Duff, K., J. Paulsen, J. Mills, L. Beglinger, D. Moser, M. Smith, D. Langbehn, S. Stout, J. Queller, and D. Harrington (2010). Mild cognitive impairment in prediagnosed huntington disease. *Neurology 75*, 500–507.

Fay, M. P. (1999). Comparing several score tests for interval censored data. *Statistics in Medicine 18*, 273–285.

Groeneboom, P. and J. A. Wellner (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation. DMV Seminar, Band 19*. New York: Birkhäuser.

Halmos, P. (1982). *A Hilbert Space Problem Book*. New York: Springer.

Harrington, D., M. Smith, Y. Zhang, N. Carlozzi, J. Paulsen, and the PREDICT-HD Investigators of the Huntington Study Group (2012). Cognitive domains that predict time to diagnosis in prodromal huntington disease. *J. Neurol Neurosurg Psychiatry 83*, 612–619.

Hu, T., B. Nan, X. Lin, and J. Robins (2011). Time-dependent cross ratio estimation for bivariate failure times. *Biometrika 98*, 341–354.

Hu, T., Q. Zhou, and J. Sun (2017). Regression analysis of bivariate current status data under

the proportional hazards model. *Canadian Journal of Statistics 45*, 410–424.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics 24*, 540–568.

Jamshidian, M. (2004). On algorithms for restricted maximum likelihood estimation. *Computational Statistics and Data Analysis 45*, 137–157.

Jewell, P. N., M. Van Der Laan, and X. Lei (2005). Bivariate current status data with univariate monitoring times. *Biometrika 92*, 847–862.

Kim, Y., J. Lim, and D. Park (2015). Testing independence of bivariate interval-censored data using modified kendall's tau statistic. *Biometrical Journal 6*, 1131–1145.

Maathuis, M. H. (2005). Reduction algorithm for the npmle for the distribution function of bivariate interval-censored data. *Journal of Computational and Graphical Statistics 14*, 352–362.

Nan, B., X. Lin, L. Lisabeth, and S. Harlow (2006). Piecewise constant cross-ratio estimation for association of age at a marker event and age at menopause. *Journal of the American Statistical Association 101*, 65–77.

Paulsen, J., J. Long, C. Ross, D. Harrington, C. Erwin, J. Williams, H. Westervelt, H. Johnson, E. Aylward, Y. Zhang, H. Bockholt, R. Barker, and the PREDICT-HD Investigators/Coordinators of Huntington Study Group (2014). Prediction of manifest huntington's disease with clinical and imaghing measures; a prospective observation study. *Lancet Neorology 13*, 1193–1201.

REFERENCES

Petersen, R. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine 256*, 183–194.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statist Science 3*, 425–441.

Schumaker, L. (1981). *Spline Function: Basic Theory*. New York: John Wiley.

Shen, X. (1997). On methods of sieves and penalization. *Annals of Statistics 25*, 2555–2591.

Shih, J. H. and T. A. Louis (1995). Inference on the assocation parameter in copula models for bivariate survival data. *Biometrics 51*, 1384–1399.

Shih, J. H. and T. A. Louis (1996). Tests of independence for bivariate survival data. *Biometrics 52*, 1440–1449.

Song, S. (2001). *Estimation With Bivariate Interval-Censored Data*. Ph. D. thesis, University of Washington.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics 10*, 1040–1053.

Sun, J. (1996). Self-consistency estimation of distributions based on truncated and doubly censored data with applications to aids cohort studies. *Statistics in Medicine 15*, 1387–1395.

Sun, L., L. Wang, and J. Sun (2006). Estimation of the association for bivariate interval-censored failure time data. *Scandinavian Journal of Statistics 33*, 637–649.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored

and truncateddata. *Journal of the Royal Statistical Society, Series B 38*, 290–295.

Walker, F. (2007). Huntington's disease. *Semin Neurol 27*, 143–150.

Wang, L., C. S. McMahan, M. G. Hudgens, and Z. P. Qureshi (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics 72*, 222–231.

Wang, W. and A. A. Ding (2000). On assessing the association for bivariate current status data. *Biometrika 87*, 1199–1215.

Wen, C. C. and Y. H. Chen (2013). A frailty model approach for regression analysis of bivariate interval-censored survival data. *Statistica Sinica 23*, 383–408.

Wong, G. Y. and Q. Yu (1999). Generalized mle of a joint distribution function with multivariate interval- censored data. *Journal of Multivariate Analysis 69*, 155–166.

Wong, W. H. (1992). On asymptotic efficiency in estimation theory. *Statistical Science 2*, 47–68.

Wu, Y. and Y. Zhang (2012). Partially monotone tensor spline estimation of the joint distribution function with bivariate current status data. *Annals of Statistics 40*, 1609–1636.

Zeng, D., F. Gao, and D. Y. Lin (2017). Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika 104*, 505–525,.

Zhang, Y., L. Hua, and J. Huang (2010). A spline-based semiparametric maximum likelihood estimation for the cox model with interval-censored data. *Scandinavian Journal of Statistics 37*, 338–354.

REFERENCES

Zhang, Y., W. Liu, and Y. Zhan (2001). A nonparametric two-sample test of the failure functions with interval censoring case 2. *Biometrika 38*, 677–686.

Zhou, Q., T. Hu, and J. Sun (2017). A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association 112*, 664–672.

Department of Biostatistics and Bioinformatics

Duke University Medical Center, Durham, NC 27705

E-mail: yuan.wu@duke.edu

Department of Biostatistics, College of Public Health

University of Nebraska Medical Center, Omaha, NE 68198

E-mail: ying.zhang@unmc.edu

Department of Biostatistics

Indiana University Fairbanks School of Public Health, Indianapolis, IN 46202

E-mail: junyzhou@iu.edu