HHS Public Access

Author manuscript

Comput Stat Data Anal. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

Comput Stat Data Anal. 2018 December ; 128: 354–366. doi:10.1016/j.csda.2018.07.016.

A Gamma-frailty proportional hazards model for bivariate interval-censored data

Prabhashi W. Withana Gamage[a], **Christopher S. McMahan**[a,*], **Lianming Wang**[b], and **Wanzhu Tu**[c]

[a]Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A.

[b]Department of Statistics, University of South Carolina, SC 29208, U.S.A.

[c]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, U.S.A.

Abstract

Correlated survival data naturally arise from many clinical and epidemiological studies. For the analysis of such data, the Gamma-frailty proportional hazards (PH) model is a popular choice because the regression parameters have marginal interpretations and the statistical association between the failure times can be explicitly quantified via Kendall's tau. Despite their popularity, Gamma-frailty PH models for correlated interval-censored data have not received as much attention as analogous models for right-censored data. In this work, a Gamma-frailty PH model for bivariate interval-censored data is presented and an easy to implement expectation-maximization (EM) algorithm for model fitting is developed. The proposed model adopts a monotone spline representation for the purposes of approximating the unknown conditional cumulative baseline hazard functions, significantly reducing the number of unknown parameters while retaining modeling flexibility. The EM algorithm was derived from a data augmentation procedure involving latent Poisson random variables. Extensive numerical studies illustrate that the proposed method can provide reliable estimation and valid inference, and is moreover robust to the misspecification of the frailty distribution. To further illustrate its use, the proposed method is used to analyze data from an epidemiological study of sexually transmitted infections.

Keywords

EM algorithm; Gamma-frailty; interval-censored data; monotone splines; multivariate regression; Poisson latent variables; proportional hazards model; survival analysis

*Corresponding author mcmaha2@clemson.edu (Christopher S. McMahan).

1. Introduction

Interval-censored data frequently arise from clinical and epidemiological studies, where outcome events are periodically assessed. In studies of sexually transmitted infections (STIs), for example, participants are often followed prospectively with predetermined testing schedules. As a result, the precise timing of infection acquisition is generally unavailable, except for the rare situations where tests are prompted by emergence of symptoms. Interval-censored data are particularly common in investigation of STIs with no or mild symptoms. For example, the motivating example considered herein involves a cohort study of young women aimed at assessing the association between certain risk factors and the contraction of STIs. In particular, this study considers *Chlamydia* t*rachomatis* and *Trichomonas vaginalis*, two organisms that cause clinical diseases of chlamydia and trichomoniasis, respectively. Moreover, individuals infected with C. *trachomatis* and T. *vaginalis* can often be asymptomatic, thus preventing knowledge of the exact acquisition time. Herein, a joint modeling approach to accommodate the known synergy between these two pathogens (Workowski and Bolan, 2015) is developed. The primary objectives of this analysis are to estimate the organism-specific survival functions, and quantify the associations between participant characteristics and risks of STI acquisition.

For correlated survival times, there are two basic modeling approaches; i.e., marginal or frailty modeling. The marginal approach specifies a marginal model for each failure time, adopts a working independence assumption in the likelihood construction, obtains point estimates of the regression parameters under this assumption, and then uses the so-called sandwich estimator to obtain standard error estimates (Wei et al., 1989). Various marginal models have been proposed along the lines of this general approach for multivariate interval-censored data; e.g., the proportional hazards (PH) model (Goggins and Finkelstein, 2000; Kim and Xue, 2002), the proportional odds (PO) model (Chen et al., 2007), the additive hazards model (Tong et al., 2008), the linear transformation model (Chen et al., 2013), and the additive transformation model (Shen, 2015). Moreover, a goodness-of-fit test for assessing the appropriateness of the marginal Cox model for multivariate interval-censored data was proposed by Wang et al. (2006). Even though the marginal approach provides robust inference, it does not adequately account for the correlation that naturally exists between the multiple failure times.

In contrast, frailty models directly acknowledge the correlation structure and introduce frailty terms in order to model the dependence between multiple responses. For this reason frailty modeling has become quite popular in survival analysis (Hougaard, 2000; Ibrahim et al., 2008; Wienke, 2012). For analyzing multivariate case 1 interval-censored data (i.e., current status data), several frailty models have been previously proposed; e.g., a probit model with normal frailty (Dunson and Dinse, 2002), a PH model with a normal frailty (Chen et al., 2009), and a PO model with a gamma-frailty (Lin and Wang, 2011). Extending to multivariate general interval-censored data, Komarek and Lessaffre (2007) proposed a frailty accelerated failure time model and Zuma (2007) explored the Gamma-frailty Weibull model.

For the analysis of correlated survival data, the Gamma-frailty proportional hazards (PH) model has proven to be a popular choice among practitioners. One advantage of this model is that the statistical association between the failure times can be explicitly (in closed-form) quantified via Kendall's τ. Research based on the Gamma-frailty PH model for multivariate right-censored data include Klein (1992), Andersen et al. (1997), Rondeau et al. (2003), Cui and Sun (2004), and Yin and Ibrahim (2005) among many others. Related work on multivariate or clustered current status data include Chang et al. (2007), Hens et al. (2009), Wen and Chen (2011), and Wang et al. (2015). In contrast, very few works have considered extending the Gamma-frailty PH model to allow for the analysis of multivariate interval-censored data, within the context studied here. For analyzing clustered interval-censored data, Lam et al. (2010) proposes a multiple imputation approach under the Gamma-frailty PH model. Similarly, Henschel et al. (2009) and Yavuz and Lambert (2016) propose frailty models for clustered interval-censored data within a Bayesian framework. To our knowledge, the work most closely related to that presented here is of Wen and Chen (2013). These authors developed an algorithm which could be used to maximize the full likelihood based on the Gamma-frailty PH model and established the asymptotic properties of their proposed estimator. However, the proposed algorithm is rather arduous to implement, even for experts in the area, and software is not readily available. In particular, the algorithm involves iteratively updating the regression parameters and the frailty variance parameter through a Newton-Raphson algorithm and solving self-consistency equations for the conditional cumulative baseline hazard functions.

Seeking to generalize Wang et al. (2015), this paper focuses on developing methods for analyzing correlated bivariate interval-censored data under the Gamma-frailty PH model. In the proposed model formulation, a monotone spline representation (Ramsay, 1988) is use to approximate the unknown conditional cumulative baseline hazard functions, thus greatly reducing the number of unknown parameters while retaining a great deal of modeling flexibility. To complete model fitting, an expectation-maximization (EM) algorithm is developed through a carefully structured data augmentation scheme involving latent Poisson random variables. This scheme leads to both straightforward parameter updates in the M-step as well as closed-form expectations in the E-step. These features make the algorithm easy to implement and computationally effcient. Moreover, through an extensive Monte Carlo simulation study, the proposed approach is shown to provide reliable estimates of all model parameters as well as valid inference, and further, is robust to the misspecification of the frailty distribution. As a companion to this work, a set of functions (coded in R) which implement all aspects of the proposed methodology have been developed and are being added to the next release of the ICsurv package, which is freely available from the CRAN (i.e., http://cran.us.r-project.org/).

The remainder of this article is organized as follows. In Section 2, the details of the proposed model and approach are presented, including but not limited to the use of monotone splines, the data augmentation steps, and the derivation of the EM algorithm. In Section 3, the results of an extensive simulation study designed to evaluate the finite sample performance of the proposed approach are provided. Section 4 provides the results of the analysis of the motivating data application; i.e., the STI data collected as a part of the Young Women's Project. Section 5 concludes with a summary discussion.

2. Model and Methodology

Let T_1 and T_2 denote the two unobserved failure/event times of interest; e.g., the time at which a patient becomes infected with C. *trachomatis* or *T. vaginalis*, respectively. To jointly model these two failure times, a Gammafrailty proportional hazards model is considered; i.e., as in Wang et al. (2015) it is assumed that the conditional cumulative hazard function for T_j, given the frailty η, is given by

$$\Lambda_j(t|\mathrm{x},\eta) = \Lambda_{0j}(t)\exp\left(\mathrm{x}'\beta_j\right)\eta, \text{for} j = 1, 2, \quad (1)$$

where x is a $(p \times 1)$-dimensional vector of covariates, β_j is the corresponding vector of regression coeffcients, $\Lambda_{0j}(\cdot)$ is the conditional cumulative baseline hazard function. Owing to the models name, the frailty (i.e., random effect) is assumed to follow a gamma distribution, whose shape and rate parameters are both ν. As is common in the literature, it is also assumed that the two failure times are conditionally independent given the frailty. It is worthwhile to point out that in order for the model to be proper $\Lambda_{0j}(\cdot)$ should be an unbounded non decreasing function with $\Lambda_{0j}(0) = 0$

By integrating over the frailty, one may ascertain that T_j marginally follows a generalized odds-rate hazards (GORH) model; i.e., the marginal survival functions for T_j is given by

$$S_j(t|\mathrm{x}) = P(T_j > t|\mathrm{x}) = \left\{1 + \nu^{-1}\Lambda_{0j}(t)\exp\left(\mathrm{x}'\beta_j\right)\right\}^{-\nu}, \text{for} j = 1, 2. \quad (2)$$

The GORH class of survival regression models is a broad family, which holds the PH and PO models as special cases. In particular, allowing $\nu \to \infty$ in (2) results in obtaining the usual PH model, while setting $\nu = 1$ provides the PO model. Noting this relationship leads to three interesting aspects of the proposed model. First, through the estimation of ν, the proposed approach is essentially identifying the best model among the GORH class for the observed data, and thus the regression coeffcients (i.e., the β_j) can be interpreted under that model as the marginal covariate effects. Secondly, a measure of association between the failure times in the form of Kendall's τ is available in closed-form and is given by $\tau = (1+2\nu)^{-1}$; for further details and discussion see Wang et al. (2015). Lastly, this realization allows for the direct assessment of the efficiency gains which can be obtained by *j*ointly modeling the failure times in contrast to modeling them marginally through the use of comparable methods; e.g., see the approach of Zhou et al. (2017).

2.1. Monotone Splines for modeling $\Lambda_{0j}(\cdot)$

The unknown parameters in the Gamma-frailty PH model involve the regression parameters β_j, the frailty variance parameter ν, and the cumulative baseline hazard function $\Lambda_{0j}(\cdot)$, for $j = 1, 2$. One could specify a functional form for $\Lambda_{0j}(\cdot)$, but proceeding in this fashion often

leads to model misspecification. Thus, in this work $\Lambda_{0j}(\cdot)$ is regarded as an unknown function and therefore represents an infinite dimensional parameter. Following the works of Wang et al. (2015), Lin and Wang (2010), Wang and Dunson (2011), Cai et al. (2011), McMahan et al. (2013), and Wang et al. (2016), the proposed approach approximates $\Lambda_{0j}(\cdot)$ through the use of the monotone regression splines of Ramsay (1988); i.e.,

$$\Lambda_{0j}(t) = \sum_{l=1}^{k_j} \gamma_{jl} I_{jl}(t), \quad (3)$$

where $I_{jl}(\cdot)$ is a monotonically increasing spline basis function, γ_{jl} is an unknown spline coefficient. To insure that $\Lambda_{0j}(\cdot)$ is a nondecreasing function, γ_{jl} is constrained to be nonnegative; i.e., $\gamma_{jl} \geq 0$, for $l = 1, \ldots, k_j$ and $j = 1, 2$. For ease of exposition, define $\gamma_j = \left(\gamma_{j1}, \ldots, \gamma_{jk_j}\right)'$.

The k_j spline basis functions considered in (3) are piecewise polynomial functions, which are fully determined by selecting a knot set, consisting of m_j points placed throughout the time domain of interest, and the degree of the polynomials (say degree$_j$), where $k_j = m_j$ +degree$_j - 2$; for further discussion see Ramsay (1988). The shape of the basis splines are predominantly determined by the placement of the knots while the degree controls the smoothness (Ramsay, 1988). For example, specifying the degree to be one, two or three corresponds to using linear, quadratic or cubic polynomials, respectively. In general, it has been suggested that specifying the degree of the polynomial basis functions to be either two or three results in adequate smoothness; e.g., see the discussion provided in McMahan et al. (2013) and Wang et al. (2016). In contrast, for modeling purposes, the selection of the number and placement of the knots plays a more important role when compared to choosing the degree, thus it is suggested that the strategies discussed in McMahan et al. (2013) and Wang et al. (2016) be adhered to when addressing this topic. In particular, these authors suggest that several knot sequences be used to complete model fitting, with model selection criterion (e.g., Akaike's information criterion or the Bayesian information criterion) being employed to determine the "best" model.

2.2. Observed data likelihood

The remainder of this work is directed towards developing and evaluating an approach to fit the model depicted in (1) to bivariate interval-censored data. Interval-censored data commonly arise in studies in which the failure/event time of interest is not directly observed but is rather known to have occurred during a time interval formed based on observation/screening times. To further elucidate, let L_j and R_j, with $L_j < R_j$, denote the two observation times which form the interval that contains T_j. Thus, if $L_j = 0$ the failure time is left-censored, if $R_j = \infty$ the failure time is right-censored, and the failure time is interval-censored otherwise. For notational convenience, let δ_{j1}, δ_{j2}, and δ_{j3} be censoring indicators

denoting left-, interval-, and right-censoring, respectively, for event j; i.e., $\delta_{j1} = I(L_j = 0)$, $\delta_{j3} = I(R_j = \infty)$, and $\delta_{j2} = 1 - \delta_{j1} - \delta_{j3}$.

Now consider a study in which bivariate interval-censored data are collected on n independent individuals; i.e., the observed data, which is given by $\mathscr{D} = \{(L_{ij}, R_{ij}, x_i, \delta_{ij1}, \delta_{ij2}, \delta_{ij3}); j = 1, 2; i = 1, 2, \ldots, n\}$, represents n independent realization of $\{(\mathrm{L}_j, \mathrm{R}_j, \mathrm{x}, \delta_{j1}, \delta_{j2}, \delta_{j3}); j = 1, 2\}$. In this case, the observed data likelihood can be expressed as

$$L(\theta) = \prod_{i=1}^{n} \int g(\eta_i | \nu) \left[\prod_{j=1}^{2} \left\{ F_j(R_{ij} | \mathrm{x}_i, \eta_i) \right\}^{\delta_{ij1}} \left\{ F_j(R_{ij} | \mathrm{x}_i, \eta_i) - F_j(L_{ij} | \mathrm{x}_i, \eta_i) \right\}^{\delta_{ij2}} \left\{ 1 - F_j(L_{ij} | \mathrm{x}_i, \eta_i) \right\}^{\delta_{ij3}} \right] d\eta_i, \tag{4}$$

where $\theta = (\beta_1', \beta_2', \gamma_1', \gamma_2', \nu)'$ is the vector of unknown parameters, $g(\cdot | \nu)$ denotes the probability density func-tion for the gamma distribution whose shape and rate parameters are both ν, and $F_j(t|\mathrm{x}, \eta)$ is the conditional cumulative distribution function of the jth failure time, given covariates x and frailty η, which is given by

$$F_j(t | \mathrm{x}, \eta) = 1 - \exp\left\{ - \Lambda_{0j}(t) \exp\left(\mathrm{x}' \beta_j \right) \eta \right\}, \text{for } j = 1, 2.$$

Note, in order to derive (4), it is assumed that the covariates are time independent and that the failure and censoring times are conditionally independent, given the covariate information. These assumptions are common among the survival literature; e.g., see Liu and Shen (2009) and Zhang and Sun (2010) and the references therein. Moreover, note that if the observed data consisted of only left- and right-censored observations (i.e., current status data) then (4) reduces to equation (3) in Wang et al. (2015).

By integrating over the gamma-frailty parameters (i.e., the η_i) one can obtain a closed-form expression for the observed data likelihood. Using this expression, it is natural to attempt to estimate the unknown parameters of the model via maximum likelihood estimation; i.e., the maximum likelihood estimator (MLE) can be obtained as $\hat{\theta} = \mathrm{argmax}_{\theta} L(\theta)$. To this end, numerical optimization techniques could be employed, but proceeding in this fashion often leads to several problems for the considered model; e.g., these techniques often converge to local extrema or experience numerical instabilities and terminate due to numerical error. In order to obviate these potential pitfalls and computational complexities, in Section 2.3 an EM algorithm is developed for the purposes of obtaining the MLE of $\boldsymbol{\theta}$.

2.3. EM algorithm

In order to facilitate the development of the proposed EM algorithm, a series of three data augmentation steps are considered. As in Wang et al. (2015), the first step of the data augmentation procedure involves introducing the individual frailties as latent random variables. Proceeding in this fashion leads to the following augmented data likelihood

$$L_1(\theta) = \prod_{i=1}^{n} g(\eta_i|\nu) \tag{5}$$
$$)\prod_{j=1}^{2}\left\{F_j\left(R_{ij}\middle|x_i,\eta_i\right)\right\}^{\delta_{ij1}}\left\{F_j\left(R_{ij}\middle|x_i,\eta_i\right)-F_j\left(L_{ij}\middle|x_i,\eta_i\right)\right\}^{\delta_{ij2}}\left\{1-F_j\left(L_{ij}\middle|x_i,\eta_i\right)\right\}^{\delta_{ij3}}.$$

Notice, by integrating (5) over the frailty terms one will obtain the observed data likelihood depicted in (4). In contrast to the data augmentation procedure of Wang et al. (2015), the second step relates the censoring indicators to latent Poisson random variables by introducing Z_{ij} and W_{ij} such that $\delta_{ij1} = I\left(Z_{ij} > 0\right), \delta_{ij2} = I\left(Z_{ij} = 0, W_{ij} > 0\right)$, and $\delta_{ij3} = I\left(Z_{ij} = 0, W_{ij} = 0\right)$, where $Z_{ij}\left|\eta_i \sim \text{Poisson}\left\{\Lambda_{0j}(t_{ij1})\exp(x_i'\beta_j)\eta_i\right\}\right.$ and $W_{ij}\left|\eta_i \sim \text{Poisson}[\left\{\Lambda_{0j}(t_{ij2}) - \Lambda_{0j}(t_{ij1})\right\}\right.\exp(x_i'\beta_j)\eta_i]$, with $t_{ij1} = R_{ij}I\left(\delta_{ij1} = 1\right) + L_{ij}I\left(\delta_{ij1} = 0\right)$ and $t_{ij2} = R_{ij}I\left(\delta_{ij2} = 1\right) + L_{ij}I\left(\delta_{ij3} = 1\right)$ Note, W_{ij} is introduced only if the failure time (i.e., T_{ij}) is interval- or right-censored, while Z_{ij} is introduced regardless of the censoring status. This additional data augmentation layer leads to the following augmented data likelihood

$$L_2(\theta) = \prod_{i=1}^{n} g(\eta_i|\nu)\prod_{j=1}^{2} Pz_{ij}\left(Z_{ij}\right)P_{W_{ij}}\left(W_{ij}\right)^{\delta_{ij2}+\delta_{ij3}}C_{ij}, \tag{6}$$

where $C_{ij} = \delta_{ij1}I(Z_{ij} > 0) + \delta_{ij2}I(Z_{ij} = 0, W_{ij} > 0) + \delta_{ij3}I(Z_{ij} = 0, W_{ij} = 0)$ and $P_Z(\cdot)$ is the probability mass function of the random variable Z. Again notice that, by integrating (6) over the latent Poisson random variables one will obtain (5). The final step exploits the monotone spline representation of $\Lambda_{0j}(\cdot)$, and decomposes Z_{ij} and W_{ij} as $Z_{ij} = \sum_{l=1}^{k_j} Z_{ijl}$ and $W_{ij} = \sum_{l=1}^{k_j} W_{ijl}$, respectively, where $Z_{ijl}|\eta_i \overset{ind.}{\sim} \text{Poisson}\left\{\gamma_{jl}I_{jl}\left(t_{ij1}\right)\exp\left(x'_i\beta_j\right)\eta_i\right\}$ and $W_{ijl}|\eta_i \overset{ind.}{\sim} \text{Poisson}\left[\gamma_{jl}\left\{I_{jl}\left(t_{ij2}\right) - I_{jl}\left(t_{ij1}\right)\right\}\exp\left(x'_i\beta_j\right)\eta_i\right]$ This last data augmentation step result in the following augmented data likelihood

$$L_C(\theta) = \prod_{i=1}^{n} g(\eta_i|\nu)\prod_{j=1}^{2}\prod_{l=1}^{k_j} P_{Z_{ijl}}\left(Z_{ijl}\right)I\left(Z_{ijl} = Z_{ij\cdot}\right)\left\{P_{W_{ijl}}\left(W_{ijl}\right)I\left(W_{ij} = W_{ij\cdot}\right)\right\}^{\delta_{ij2}+\delta_{ij3}} C_{ij}, \tag{7}$$

where $Z_{ij\cdot} = \sum_{l=1}^{k_j} Z_{ijl}$ and $W_{ij\cdot} = \sum_{l=1}^{k_j} W_{ijl}$. Again ,by integrating (7) over the latent Poisson random variables introduced in this step (i.e., the Z_{ij} and W_{ij}) one obtains (6). For the purposes of deriving the EM algorithm (7) will be viewed as the complete data

likelihood, with the aforementioned latent variables being treated as missing data. It is worthwhile to point out that the final data augmentation step is introduced so that closed-form updates of the spline coefficients can be obtained in the M-step of the algorithm.

In general, the EM algorithm consists of two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, one takes the expectation of the logarithm of (7) with respect to all of the latent variables introduced through the aforementioned data augmentation steps, conditional on the current parameter value $\theta^{(d)} = \left(\beta_1^{(d)'}, \beta_2^{(d)'}, \gamma_1^{(d)'}, \gamma_2^{(d)'}, \nu^{(d)}\right)'$ and the observed data $\mathscr{D}$. This process results in obtaining what is referred to as the $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ function; i.e., $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) = E[\log\{L_c(\boldsymbol{\theta})\}|\mathscr{D}, \boldsymbol{\theta}^{(d)}]$. The M-step then finds $\boldsymbol{\theta}^{(d+1)} = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$. These two steps are then iterated in turn until convergence. The details involved in completing these two steps are now provided. First note, as in Wang et al. (2015), the E-step yields

$$Q\left(\theta, \theta^{(d)}\right) = H_1\left(\theta, \theta^{(d)}\right) + H_2\left(\theta, \theta^{(d)}\right) + H_3\left(\theta^{(d)}\right),$$

Where

$$H_1\left(\theta, \theta^{(d)}\right) = n\nu\log(\nu) - n\log\{\,\Gamma\,(\nu)\} + \nu\sum_{i=1}^{n}\left[E\left\{\log\left(\eta_i\right)\right\} + E\left(\eta_i\right)\right], \quad (8)$$

$$H_2\left(\theta, \theta^{(d)}\right) = \sum_{i=1}^{n}\sum_{j=1}^{2}\sum_{l=1}^{k_j}\left[\left\{E\left(Z_{ijl}\right) + \delta_{ij2}E\left(W_{ijl}\right)\right\}\left\{\log\left(\gamma_{jl}\right) + \mathrm{x}_i'\beta_j\right\}\right. \quad (9)$$

$$\left.-\gamma_{jl}\left\{\left(\delta_{ij1} + \delta_{ij2}\right)I_{jl}\left(R_{ij}\right) + \delta_{ij3}I_{jl}\left(L_{ij}\right)\right\}\exp\left(\mathrm{x}_i'\beta_j\right)E\left(\eta_i\right)\right],$$

and $H_3(\boldsymbol{\theta}^{(d)})$ is a function of $\boldsymbol{\theta}^{(d)}$ but is free of $\boldsymbol{\theta}$. Note, a simplifying step has been taken to reach (9) which involves dropping $\delta_{ij3}E(W_{ijl})$ since it is always equal to zero; i.e., $E(W_{ijl}) = 0$ when $\delta_{ij3} = 1$ and the product is obviously equal to zero when $\delta_{ij3} = 0$. At this point several comments are warranted. First and foremost, the dependencies in the conditional expectations depicted in (8) and (9) are suppressed for ease of exposition; i.e., $E(\cdot) \equiv E(\cdot|\mathscr{D}, \boldsymbol{\theta}^{(d)})$ from henceforth. Moreover, these expectations exist in closed-form and are provided in Web Appendix A of the Supplementary Material. Second, structurally the expressions provided in (8) and (9) are very similar to their counterparts in Wang et al. (2015), with subtle yet very stark differences. These differences primarily arise in the form of the expectations and the structure of (9) and allow the proposed approach to accommodate interval-censored observations, unlike this existing technique.

To complete the M-step of the algorithm, one must obtain $\boldsymbol{\theta}^{(d+1)}$. First, note that maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ with respect to ν is tantamount to maximizing (8) with respect to the same. Thus, consider the partial derivative of (8) with respect to ν, which is given by

$$\frac{\partial H_1\left(\theta,\theta^{(d)}\right)}{\partial \nu} = n\log(\nu) + n - n\psi(\nu) + \sum_{i=1}^{n}\left[E\left\{\log\left(\eta_i\right)\right\} - E\left(\eta_i\right)\right],$$

where $\psi(\nu)$ is the digamma function. Thus, solving $\partial H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})/\partial \nu = 0$ for ν, one obtains the value of $\nu^{(d+1)}$, and this step can easily be completed using standard root finding algorithms; e.g., uniroot in R. Similarly, to find $\beta_j^{(d+1)}$ and $\gamma_{jl}^{(d+1)}$, one needs only to maximize $H_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ with respect to $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$. To this end, consider the partial derivatives of $H_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ with respect to $\boldsymbol{\gamma}_{jl}$ which is given by

$$\frac{\partial H_2\left(\theta,\theta^{(d)}\right)}{\partial \gamma_{jl}} = \sum_{i=1}^{n}$$
$$\left[\gamma_{jl}^{-1}\left\{E\left(Z_{ijl}\right) + \delta_{ij2}E\left(W_{ijl}\right)\right\} - \left\{\left(\delta_{ijl} + \delta_{ij2}\right)I_{jl}\left(R_{ij}\right) + \delta_{ij3}I_{jl}\left(L_{ij}\right)\right\}\exp\left(\mathrm{x}_i'\beta_j\right)E\left(\eta_i\right)\right],$$

for $l=1,\ldots,k_j$ and $j=1, 2$. Setting this expression equal to zero and solving result in obtatining

$$\tilde{\gamma}_{jl}\left(\beta_j\right) = \frac{\sum_{i=1}^{n} E\left(Z_{ijl}\right) + \delta_{ij2}E\left(W_{ijl}\right)}{\sum_{i=1}^{n}\left\{\left(\delta_{ij1} + \delta_{ij2}\right)I_{jl}\left(R_{ij}\right) + \delta_{ij3}I_{jl}\left(L_{ij}\right)\right\}\exp\left(\mathrm{x}_i'\beta_j\right)E\left(\eta_i\right)},$$

as the solution for $l = 1,\ldots,k_j$ and $j = 1, 2$. It is worthwhile to note that $\tilde{\gamma}_{jl}(\beta j)$ depends on the value of $\boldsymbol{\beta}_j$ and further that $\tilde{\gamma}_{jl}(\beta j) \geq 0$ since all quantities in the ratio are greater than or equal to zero. That is, this quantity naturally adheres to the constraint necessary to ensure the monotonicity of $\Lambda_{0j}(\cdot)$, for all values of the regression coefficient. Now consider the system of equations that arise from taking the partial derivatives of $H_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ with respect to β_j and setting it equal to zero; i.e.,

$$\sum_{i=1}^{n}\left[\left\{E\left(Z_{ij}\right) + \delta_{ij2}W\left(W_{ij}\right)\right\} - \left\{\left(\delta_{ij1} - \delta_{ij2}\right)\Lambda_{0j}\left(R_{ij}\right) + \delta_{ij3}\Lambda_{0j}\left(L_{ij}\right)\right\}\exp\left(\mathrm{x}_i'\beta_j\right)E\left(\eta_i\right)\right]\mathrm{x}_i' = 0\,. \tag{10}$$

Replacing, $\boldsymbol{\gamma}_{jl}$ by $\tilde{\gamma}_{jl}\left(\beta_j\right)$ in (10) and solving for $\boldsymbol{\beta}_j$ results in obtaining $\beta_j^{(d+1)}$, and thus $\gamma_{jl}^{(d+1)} = \tilde{\gamma}_{jl}\left(\beta_j^{(d+1)}\right)$. Following the work of Wang et al. (2016), it is relatively easy to establish that the updated regression and spline coefficients are the unique maximizers of $H_2(\theta, \theta^{(d)})$. Thus, after setting $d = 0$ and initializing $\boldsymbol{\theta}^{(d)}$, the proposed EM algorithm repeats the following steps until a convergence criterion has been met.

1. Obtain $\nu^{(d+1)}$ as the solution to $\sum_{i=1}^{n}\left[E\left\{\log\left(\eta_i\right)\right\} - E\left(\eta_i\right)\right] = n\psi(\nu) - n\log(\nu) - n$.

2. Obtain $\beta_j^{(d+1)}$, for $j = 1, 2$, as the solution to the following system of p equations

$$\sum_{i=1}^{n}\left\{E\left(Z_{ij}\right)+\delta_{ij2}E\left(W_{ij}\right)\right\}\mathrm{x}_i' = \sum_{i=1}^{n}\sum_{l=1}^{k_j}\tilde{\gamma}_{jl}\left(\beta_j\right)\left\{\left(\delta_{ij1}+\delta_{ij2}\right)I_{jl}\left(R_{ij}\right)+\delta_{ij3}I_{jl}\left(L_{ij}\right)\right\}\exp\left(\mathrm{x}_i'\beta_j\right)E\left(\eta_i\right)\mathrm{x}_i'.$$

3. Calculate $\gamma_{jl}^{(d+1)} = \tilde{\gamma}_{jl}\left(\beta_j^{(d+1)}\right)$, for $l=1,\ldots,k_j$ and $j = 1, 2$.

4. Set $d = d + 1$, and return to 1.

At the point of convergence of the EM algorithm, the MLE of $\boldsymbol{\theta}$ is obtained as $\hat{\theta} = \left(\hat{\beta}_1', \hat{\beta}_2', \hat{\gamma}_1', \hat{\gamma}_2', \hat{\nu}\right) = \theta^{(d)}$.

2.4. Variance estimation

In order to conduct large sample inference, it is suggested that the asymptotic covariance matrix be estimated via the outer product of gradients estimator, which is given by

$$\hat{V}\left(\hat{\theta}\right) = \left[\frac{1}{n}\sum_{i=1}^{n} i_i\left(\hat{\theta}\right)i_i'\left(\hat{\theta}\right)\right]^{-1},$$

Where $i_i\left(\hat{\theta}\right) = \partial l_i(\theta)/\partial\theta\big|_{\theta=\hat{\theta}}$ and $l_i(\theta)$ is the log-likelihood contribution of the ith individual, which can be expressed in terms of the marginal and joint survival functions; for further details see Web Appendix B of the Supplementary Material. Other more traditional estimators were considered; e.g., Louis's method (Louis, 1982) and the usual observed Fisher information. The details required to implement the former were found to be rather complex, while the latter provided standard error estimates that were at times less than satisfactory.

3. Simulation Study

In order to investigate the finite sample performance of the proposed methodology, the following simulation study was conducted. The true distribution of the failure time T_j, for $j = 1, 2$, was specified to be

$$F_j(t|\mathrm{x},\eta) = 1 - \exp\left\{-\Lambda_{0j}(t)\exp\left(x_1\beta_{j1}+x_2\beta_{j2}\right)\eta\right\}$$

where $\Lambda_{0j}(t) = \log\left(t^2+1\right)$, $\mathrm{x}=\left(x_1, x_2\right)'$, $x_1 \sim$ Bernoulli(0.5), $x_2 \sim \mathrm{N}(0, 0.5^2)$, $\eta \sim$ Gamma(ν, ν), where $\nu \in \{0.25, 1, 4\}$. These values of ν emit a small ($\nu = 4$), moderate ($\nu = 1$), and large ($\nu = 0.25$) association between the two failure times. The regression coefficients (i.e., β_{j1} and β_{j2}) were specified such that, $\beta_{11} = \beta_{21}$ and $\beta_{12} = \beta_{22}$, with each taking values −0.5, 0, and 0.5. These specifications result in nine different configurations of the regression parameters.

In order to simulate the observed data, the failure time T_j was first determined by solving $F_j(t|x, \eta) = u$, where $u \sim$ Uniform(0, 1). Observation times were generated through an independent observational process having support on the interval (0,10). The number of observation times were determined as one plus a Poisson random variable having mean parameter three. This assures that each individual has at least one observation time, but allows the number of observation times to vary across individuals. The waiting times between adjacent observations were generated according to an exponential distribution with mean one. Thus, L_j and R_j were determined by examining which of the two observation times bounded the failure time, with the convention that if T_j was smaller (greater) than the smallest (largest) observation time then $L_j = 0$ ($R_j = \infty$).

The aforementioned process was used to randomly generate 500 datasets, each consisting of n = 500 observations, for all of the considered parameter configurations. The proposed EM algorithm was then used to analyze each of the resulting data sets. To implement the algorithm, a separate monotone spline representation was used for each of the failure times, with these specifications being based on the set of available observation times. To provide several configurations, the degree for both spline functions were set to be equal and took values of two and three. A knot set consisting of $m_1 = m_2 = m \in \{3, 4, 5\}$ knots were considered. These specifications lead to a total of six different spline representations which were used to analyze each data set. In each case, the boundary knots were placed at the minimum and maximum of the observed finite time points and the interior knots were placed at evenly space quantiles of the finite nonzero time points; e.g., when m = 5 the three interior knots were placed at the first, second, and the third quartiles. The starting value was specified to be $\theta^{(0)} = \left(\beta_1^{(0)\prime}, \beta_2^{(0)\prime}, \gamma_1^{(0)\prime}, \gamma_2^{(0)\prime}, \nu^{(0)}\right) = \left(0'_2, 0'_2, 1'_{k_1}, 1'_{k_2}, 1\right)$, where $\mathbf{0}_q(\mathbf{1}_q)$ is a ($q \times 1$)-dimensional vector of zeros (ones). Convergence was declared when the maximum absolute difference between consecutive parameter updates was less than the specified tolerance of 0.001.

In order to provide a comparison between the proposed method and existing techniques, two competing approaches were considered. The first technique, which from henceforth will be referred to as the univariate approach, considered modeling each of the failure times separately using the GORH model. To accomplish this, the ICGOR package in R was used to fit the marginal GORH model depicted in (2); for further details see Zhou et al. (2017). This package implements an EM algorithm for the purposes of estimating both the regression and spline coefficients for a fixed value of ν, with ν being estimated through the implementation of a grid search across a sequence of feasible values. The method implemented by the ICGOR package also makes use of the monotone spline representation depicted in (3) to approximate the unknown cumulative baseline hazard function. Thus, to provide a fair comparison, the degree and number of interior knots were specified to be the same as the proposed approach. The ICGOR package also provides standard error estimates by an appeal to Louis's method (Louis, 1982). It is important to note that this approach does not acknowledge the fact/potential that the failure times are related. To acknowledge dependence, a common approach involves fitting both of the marginal models and then correcting the standard errors via a joint sandwich estimator of the asymptotic covariance matrix, for further details see Freedman (2006). This approach was also implemented and is

referred to as the marginal approach. Note, a feature of the univariate and marginal methods is that they render the exact same regression and spline coefficient estimates, but they provide for different standard error estimates, with the former ignoring the dependence between the failure times and the latter accounting for it.

Table 1 summarizes the estimates of the regression coefficients obtained by the proposed approach across all of the considered regression parameter configurations, when $\nu = 1$, $m = 5$, and degree is three. Web Tables 1–17 provide the analogous summary for the other considered values of ν, m, and degree. This summary includes the empirical bias, the sample standard deviation of the 500 point estimates, the average of the standard error estimates, and the empirical coverage probabilities associated with 95% confidence intervals for the regression coefficients, for each of the failure times. From these results, one will first note that the proposed approach results in estimates that exhibit little if any evidence of bias. Additionally, the sample standard deviation of the 500 point estimates obtained from the proposed approach is in agreement with the average of the standard error estimates, indicating that the outer product of gradients estimator suggested in Section 2.4 is appropriate for conducting finite sample inference. This is supported by the fact that the empirical coverage probabilities for the regression parameters are all at their nominal level. Further, from the additional results presented in Web Tables 1–17 it appears that the proposed approach is relatively robust to the specification of the spline functions. That is, no appreciable differences are apparent in these additional results.

Table 1 also summarizes the parameter estimates arising from the two competing techniques; i.e., the univariate and marginal methods. From these results one will note that the two competing techniques perform well, but differences are apparent when comparisons are made with the proposed approach. In particular, the parameter estimates obtained from the two competing techniques are in general less efficient (i.e., possess more variability) than those obtained from the proposed approach. Further, the estimates from the competing techniques also exhibit a significantly larger bias when compared to the estimates resulting from the proposed approach; e.g., the empirical bias for the univariate and marginal methods were between 2 and 13 times larger than those resulting from the proposed approach. These losses in both estimation efficiency and precision are likely attributable to two features; first, the fact that both the univariate and marginal approach ignore, during estimation, the dependence which exists between the failure times, and second, that fitting the marginal GORH model is a relatively difficult process due to the estimation of a frailty parameter; for further discussion see Zhou et al. (2017). Moreover, the empirical coverage probabilities for both of the competing techniques were rarely at their nominal level, with the univariate and marginal methods tending to under and over cover, respectively. Additionally, the sandwich estimator employed by the marginal approach appears to egregiously over estimate the standard errors for the regression coefficients in some instances, this can be seen when one compares the average standard error estimates to the medians, see Table 1. Similarly, the univariate method occasionally provided negative standard error estimates for the regression parameters, in these cases the estimates were omitted when calculating the average standard errors and empirical coverage probabilities. It is important to note, the proposed approach did not encounter these issues when used to estimate standard errors.

Table 2 summarizes the estimates of ν obtained by the proposed approach across all considered simulation configurations, when $m = 5$ and degree is three. Web Tables 18–22 provide the analogous summary for the other considered values m and degree. This summary includes the empirical bias, the sample standard deviation of the 500 point estimates, the average of the standard error estimates, and the empirical coverage probabilities associated with 95% confidence intervals for ν. For a moderate to a large association (i.e., when $\nu = 0.25$ and 1) these estimates exhibit very little evidence of bias, and the sample standard deviation and the averaged standard errors of the 500 point estimates are generally in agreement. Further, the empirical coverage probabilities are also generally at their nominal level. It is worthwhile to point out that when there is a small association between the failure times (i.e., $\nu = 4$) the estimation and inference associated with ν becomes a bit strained; i.e., the bias has the propensity to be markedly larger, there tend to be disagreement between the sample standard deviation and the averaged standard errors, and the 95% confidence intervals tend to over cover. Although, even in this case the estimation and inference associated with the regression coefficients is not negatively impacted. In some sense, this finding is not so surprising; i.e., ν essentially controls the amount of dependence between the failure times, if the dependence is weak then there is a lack of information available to estimate it.

Figure 1 summarizes the estimates of the baseline survival functions (i.e., $S_{0j}(t) = S_j(t|\mathrm{x} = 0_p)$) for failure time 1, across all considered regression parameter configurations when $\nu = 1$, $m = 5$, and degree is three. The analogous figures for the other considered simulation configurations are provided in Web Figures 1–35. This figure presents plots of the average estimate along with curves representing the pointwise 2.5th and 97.5th percentiles of the estimates. Also provided are curves representing the true baseline survival functions. These figures illustrate that the proposed approach can accurately estimate the baseline survival functions of the two failure times, which is tantamount to well estimating the conditional cumulative baseline hazard function.

In synopsis, this simulation study has served to illustrate that the proposed methodology is capable of accurately estimating the unknown model parameters and renders reliable inference. Moreover, this study has illustrated that the proposed method is superior when compared to the univariate and marginal methods. Thus, these findings tend to suggest that the proposed approach would be preferable for the purposes of analyzing dependent bivariate interval-censored data when compared to the two considered existing techniques.

3.1. Simulation Study II

An additional robustness study was conducted in order to ascertain the impact of misspecifying the frailty distribution. This study considers the exact same data generating process described above with the exception that the frailty distribution was misspecified. In particular, three such frailty distributions were considered:

$$f_1(\eta) = 0.25\mathrm{LN}(-1.20, 1.85) + 0.50\mathrm{LN}(-0.90, 0.56) + 0.25\mathrm{LN}(0.60, 0.23),$$

$$f_2(\eta) = 0.20\mathrm{LN}(-1.20, 1.85) + 0.20\mathrm{LN}(-0.90, 0.56) + 0.60\mathrm{LN}(0.60, 0.23), \text{ and}$$

$$f_3(\eta) = 0.30\mathrm{WN}(3.00, 0.60) + 0.40\mathrm{WN}(2.50, 1.80) + 0.40\mathrm{WN}(4.50, 1.00),$$

where LN(μ, σ^2) denotes the lognormal distribution with location parameter μ and scale parameter σ and WL(κ, λ) denotes the Weibull distribution with shape parameter κ and scale parameter λ. Under each of these frailty models, 500 data sets, each consisting of n = 500 observations, were randomly generated, and the proposed method was used to analyze each in the exact same fashion as was described above, with the degree of both spline functions being set to three and $m_1 = m_2 = 5$. Table 3 summarizes the parameter estimates for this study. These results again illustrate that the proposed method performs well; i.e., bias is small, averaged standard errors and sample standard deviations are in agreement, and the empirical coverage probabilities for the regression coefficients are at their nominal. This robustness study shows that the proposed approach is not unduly impacted by the misspecification of the frailty distribution.

4. Data Application

To illustrate the use of the proposed model, data from a longitudinal study of STIs was analyzed. The study design and follow-up protocol were previously described (Tu et al., 2009; Ghosh and Tu, 2009; Tu et al., 2011; Yu et al., 2012). Briefly, young women between 14 and 17 years of age were recruited for participation in this prospective cohort study. Upon enrollment, participants completed face-to-face interviews and detailed questionnaire about their sexual behaviors, and they were tested for infections with *C. trachomatis* and *T. vaginalis*. Infected individuals were treated promptly. During the course of follow-up, participants were scheduled to be tested every three months, although the actual test dates could deviate from the testing schedule.

The current analysis focuses on the time from sexual debut till the first infection acquisition with *C. trachomatis* and *T. vaginalis*. For those who were sexually active at enrollment, the age of sexual debut was determined from the enrollment interview. For those who became sexually active during the study, the time of sexual debut was determined from follow-up interviews. The precise dates of infection acquisition were interval-censored by the two testing dates flanking the interval at which *C. trachomatis* and *T. vaginalis* were first detected. Time to infection was right-censored at the end of the study if the participant tested negative throughout the follow-up.

This analysis examines associations between STI acquisition and several participant characteristics, including the number of lifetime partners reported at the time of enrollment (x_1), self-reported age at sexual debut (x_2), and race (x_3 = 1 if African American, and x_3 = 0 otherwise). Twenty-seven participants were excluded from the analysis due to missing data. Other data discrepancies warranted exclusion of another nine participants; i.e., if a participant had reported an age of sexual debut later than infection detection. After the due

diligence steps on data quality, a subset of participants ($n = 350$) were included in the current analysis. Among these individuals, 37.1%, 30%, and 32.9% were left-, interval-, and right-censored, respectively, for *C. trachomatis* and 17.4%, 22.3%, and 60.3% were left-, interval-, and right-censored, respectively, for *T. vaginalis*.

This analysis considers relating the three available covariates to the time of STI acquisition through the proposed Gamma-frailty PH model, where each covariate is entered as a first order term. The proposed EM algorithm was used to fit the Gamma-frailty PH model to the STI data. The algorithm was implemented using a separate monotone spline representation for each of the event times, with these specifications being based on the set of available follow-up times. In particular, the degree of the splines was set to be three and a knot set consisting of two boundary and three interior knots was considered. The boundary knots were placed at the minimum and maximum of the observed finite follow-up times and the three interior knots were placed at the first, second, and the third quartiles of the finite nonzero time points. A starting value for the algorithm and convergence was determined in the exact same fashion as was described in Section 3. Further, the univariate and marginal methods were also implemented. A summary of the regression parameter estimates (and their estimated standard errors) obtained from these three techniques are presented in Table 4.

This analysis indicates that a larger numbers of lifetime partners at baseline, older age of sexual debut, and being African American were associated with an increased risk of *C. trachomatis* infection. For *T. vaginalis*, only being African American was associated with an increased risk of early infection acquisition. The univariate approach led to the same general conclusions with the exception that it did not identify self-reported age of sexual debut as being associated with the acquisition of *C. trachomatis*. In contrast, the marginal approach did not detect any significant associations, with the exception of race being related to *T. vaginalis* infection. The discrepancies observed between the approaches are likely attributable to the observations discussed in Section 3; i.e., by modeling the data jointly the proposed method provides more efficient and precise estimates, as well as more reliable inference. Moreover, the proposed method is able to quantify the association between the two event times; i.e., the proposed method estimated ν to be $\hat{\nu} = 2.0549$, which translates to a moderate degree of association ($\hat{\tau} = 0.1957$) between the first detections of C. trachomatis and T. vaginalis. To assess model adequacy, Figure 2 provides the average estimated survival functions (stratified by race) from the proposed and univariate/marginal methods. This figure also provides nonparametric estimates of the survival curves based on the Turnbull estimator (Turnbull, 1976), again stratified by race. These results tend to suggest that the proposed approach provides a good fit to these data, especially when compared to the fits from the univariate/marginal method.

5. Discussion

In this paper, a new EM algorithm was developed which can be used to fit the Gamma-frailty PH model to bivariate interval-censored data. The proposed formulation of the Gamma-frailty PH model makes use of a monotone spline representation to approximate the unknown conditional cumulative baseline hazard function. The derivation of the algorithm is

based on a three stage data augmentation procedure involving latent Poisson random variables and gamma-frailty terms. Based on these steps, all of the expectations necessary to implement the EM algorithm are provided in closed-form. Moreover, the regression and gamma-frailty variance parameters are obtained by solving a low-dimensional system of equations and the spline coefficients are updated in closed-form. The resulting EM algorithm is easy to implement, is robust to initialization, and enjoys quick convergence. Through Monte Carlo simulation studies, it has been shown that the proposed method performs well with respect to estimating the regression parameters, spline coefficients, and gamma-frailty variance parameter. The finite sample performance of the proposed approach was further illustrated by applying the method to interval-censored STI data collected on young women as a part of the Young Women's Project. In summary, the proposed method provides an accurate and reliable approach that can be used to analyze bivariate interval-censored data. To further disseminate this work, a set of functions (coded in R), along with supporting documentation, have been developed and are being added to the next release of the ICsurv package, which is freely available from the CRAN (i.e., http://cran.us.rproject.org/). Further, this software is available from the corresponding author upon request.

It is worthwhile to point out that the methodology proposed in this manuscript could be extended to account for more than two event/failure times; i.e., $J > 2$. Although, there would be several hurdles. First and foremost, by virtue of how the model comes together, this extension would provide for the same dependence structure between the multiple events times; which could be unreasonable for some applications. The second hurdle involves a combinatorial explosion in the number of terms that would need to be computed to complete the Estep; i.e., there are 3^J different failure time combinations, each producing a different expectation for each of the latent variables, moreover the number of latent variables also increases as a power of J. Even in lieu of these hurdles this extension could be an interesting topic of future research given the appropriate motivating example.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by National Institutes of Health grants AI121351, HD042404, and AA025208.

References

Andersen P, Klein J, Knudsen K, Palacios R, 1997 Estimation of variance in cox's regression model with shared gamma frailties. Biometrics 53, 1475–1484. doi:10.2307/2533513. [PubMed: 9423262]

Cai B, Lin X, Wang L, 2011 Bayesian proportional hazards model for current status data with monotone splines. Computational Statistics and Data Analysis 55, 2644–2651. doi:10.1016/j.csda.2011.03.013.

Chang I, Wen C, Wu Y, 2007 A profile likelihood theory for the correlated gamma-frailty model with current status family data. Statistica Sinica 17, 1023–1046.

Chen M, Tong X, Sun J, 2007 The proportional odds model for multivariate interval-censored failure time data. Statistics in Medicine 26, 5147–5161. doi:10.1002/sim.2907. [PubMed: 17476643]

Chen M, Tong X, Sun J, 2009 A frailty model approach for regression analysis of multivariate current status data. Statistics in Medicine 28, 3424–3436. doi:10.1002/sim.3715. [PubMed: 19739240]

Chen M, Tong X, Zhu L, 2013 A linear transformation model for multivariate interval-censored failure time data. The Canadian Journal of Statistics 41, 275–290. doi:10.1002/cjs.11177.

Cui S, Sun Y, 2004 Checking for the gamma frailty distribution under the marginal proportional hazards frailty model. Statistica Sinica 14, 249–267.

Dunson D, Dinse G, 2002 Bayesian models for multivariate current status data with informative censoring. Biometrics 58, 79–88. doi:10.1111/j.0006-341X.2002.00079.x. [PubMed: 11890330]

Freedman D, 2006 On the so-called huber sandwich estimator and robust standard errors. The American Statistician 60, 299–302. doi:10.1198/000313006X152207.

Ghosh P, Tu W, 2009 Assessing sexual attitudes and behaviors of young women: a *j*oint model with nonlinear time effects, time varying covariates, and dropouts. Journal of the American Statistical Association 104, 474–doi:10.1198/016214508000000850.

Goggins W, Finkelstein D, 2000 A proportional hazards model for multivariate interval-censored failure time data. Biometrics 56, 940–943. doi:10.1111/j.0006-341X.2000.00940.x. [PubMed: 10985240]

Hens N, Wienke A, Aerts M, Molenberghs G, 2009 The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. Statistics in Medicine 28, 2785–2800. doi:10.1002/sim.3660. [PubMed: 19591117]

Henschel V, Engel J, H¨olzel D, Mansmann U, 2009 A semiparametric bayesian proportional hazards model for interval censored data with frailty effects. BMC medical research methodology 9, 9. [PubMed: 19208234]

Hougaard P, 2000 Analysis of multivariate survival data Springer: New York.

Ibrahim J, Chen M, Sinha D, 2008 Bayesian survival analysis Springer: New York doi: 10.1002/0470011815.b2a11006.

Kim M, Xue X, 2002 The analysis of multivariate interval-censored survival data. Statistics in Medicine 21, 3715–3726. doi:10.1002/sim.1265. [PubMed: 12436466]

Klein J, 1992 Semiparametric estimation of random effects using the cox model based on the em algorithm. Biometrics 48, 795–806. doi:10.2307/2532345. [PubMed: 1420842]

Komarek A, Lessaffre E, 2007 Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution. Statistica Sinica 17, 549–569.

Lam K, Xu Y, Cheung T, 2010 A multiple imputation approach for clustered interval-censored survival data. Statistics in medicine 29, 680–693. [PubMed: 20069624]

Lin X, Wang L, 2010 A semiparametric probit model for case 2 interval-censored failure time data. Statistics in Medicine 29, 972–981. doi:10.1002/sim.3832. [PubMed: 20069532]

Lin X, Wang L, 2011 Bayesian proportional odds models for analyzing current status data: univariate, clustered, and multivariate. Communications in Statistics - Simulation and Computation 40, 1171–1181. doi:10.1080/03610918.2011.566971.

Liu H, Shen Y, 2009 A semiparametric regression cure model for interval-censored data. Journal of the American Statistical Association 104, 1168–1178. doi:10.1198/jasa.2009.tm07494. [PubMed: 20354594]

Louis T, 1982 Finding the observed information matrix when using the em algorithm. Journal of the Royal Statistical Society: Series B 44, 226–233.

McMahan C, Wang L, Tebbs J, 2013 Regression analysis for current status data using the em algorithm. Statistics in Medicine 32, 4452–4466. doi:10.1002/sim.5863. [PubMed: 23761135]

Ramsay J, 1988 Monotone regression splines in action. Statistical Science 3, 425–461. doi:10.1214/ss/1177012761.

Rondeau V, Commenges D, Joly P, 2003 Maximum penalized likelihood estimation in a gamma-frailty model. Lifetime Data Analysis 9, 139–153. doi:10.1023/A:1022978802021. [PubMed: 12735493]

Shen P, 2015 Additive transformation models for multivariate interval-censored data. Communications in Statistics-Theory and Methods 44, 1065–1079. doi:10.1080/03610926.2012.762398.

Tong X, Chen M, Sun J, 2008 Regression analysis of multivariate interval-censored failure time data with application to tumorigenicity experiments. Biometrical Journal 50, 364–374. doi:10.1002/bimj.200710418. [PubMed: 18435503]

Tu W, Batteiger B, Wiehe S, Ofner S, Pol BVD, Katz B, Orr D, Fortenberry J, 2009 Time from first intercourse to first sexually transmitted infection diagnosis among adolescent women. Archives of pediatrics & adolescent medicine 163, 1106–1111. doi:10.1001/archpediatrics.2009.203. [PubMed: 19996047]

Tu W, Ghosh P, Katz B, 2011 A stochastic model for assessing chlamydia trachomatis transmission risk by using longitudinal observational data. Journal of the Royal Statistical Society: Series A 174, 975–989. doi:10.1111/j.1467-985X.2011.00691.x.

Turnbull B, 1976 The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society: Series B (Methodological) 38, 290–295. doi:10.2307/2984980.

Wang L, Dunson D, 2011 Semiparametric bayes' proportional odds models for current status data with underreporting. Biometrics 67, 1111–1118. doi:10.1111/j.1541-0420.2010.01532.x. [PubMed: 21175554]

Wang L, McMahan C, Hudgens M, Qureshi Z, 2016 A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. Biometrics 72, 222–231. doi:10.1111/biom.12389. [PubMed: 26393917]

Wang L, Sun L, Sun J, 2006 A goodness-of-fit test for the marginal cox model for correlated interval-censored failure time data. Biometrical Journal 48, 1020–1028. doi:10.1002/bimj.200510269. [PubMed: 17240659]

Wang N, Wang L, McMahan C, 2015 Regression analysis of bivariate current status data under the gamma-frailty proportional hazards model using the em algorithm. Computational Statistics and Data Analysis 83, 140–150. doi:10.1016/j.csda.2014.10.013.

Wei L, Lin D, Weissfeld L, 1989 Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association 84, 1065–1073. doi:10.1080/01621459. 1989.10478873.

Wen C, Chen Y, 2011 Nonparametric maximum likelihood analysis of clustered current status data with the gamma-frailty cox model. Computational Statistics and Data Analysis 55, 1053–1060. doi:10.1016/j.csda.2010.08.013.

Wen C, Chen Y, 2013 A frailty model approach for regression analysis of bivariate interval-cenosred survival data. Statistica Sinica 23, 383–408. doi:10.5705/ss.2011.151.

Wienke A, 2012 Frailty models in survival analysis Chapman & Hall .

Workowski K, Bolan G, 2015 Sexually transmitted diseases treatment guidelines. Morbidity and Mortality Weekly Report (MMWR) 64, 1–137. [PubMed: 25590678]

Yavuz AÇ, Lambert P, 2016 Semi-parametric frailty model for clustered interval-censored data. Statistical Modelling, 360–391.

Yin G, Ibrahim J, 2005 A class of bayesian shared gamma frailty models with multivariate failure time data. Biometrics 61, 208–216. doi:10.1111/j.0006-341X.2005.030826.x. [PubMed: 15737095]

Yu Z, Lin X, Tu W, 2012 Semiparametric frailty models for clustered failure time data. Biometrics 68, 429–436. doi:10.1111/j.1541-0420.2011.01683.x. [PubMed: 22070739]

Zhang Z, Sun J, 2010 Interval censoring. Statistical Methods in Medical Research 19, 53–70. doi:10.1177/0962280209105023. [PubMed: 19654168]

Zhou J, Zhang J, Lu W, 2017 An expectation maximization algorithm for fitting the generalized odds-rate model to interval censored data. Statistics in medicine 36, 1157–1171. [PubMed: 28004414]

Zuma K, 2007 A bayesian analysis of correlated interval-censored data. Communications in Statistics-Theory and Methods 36, 725–730. doi:10.1080/03610920601033710.

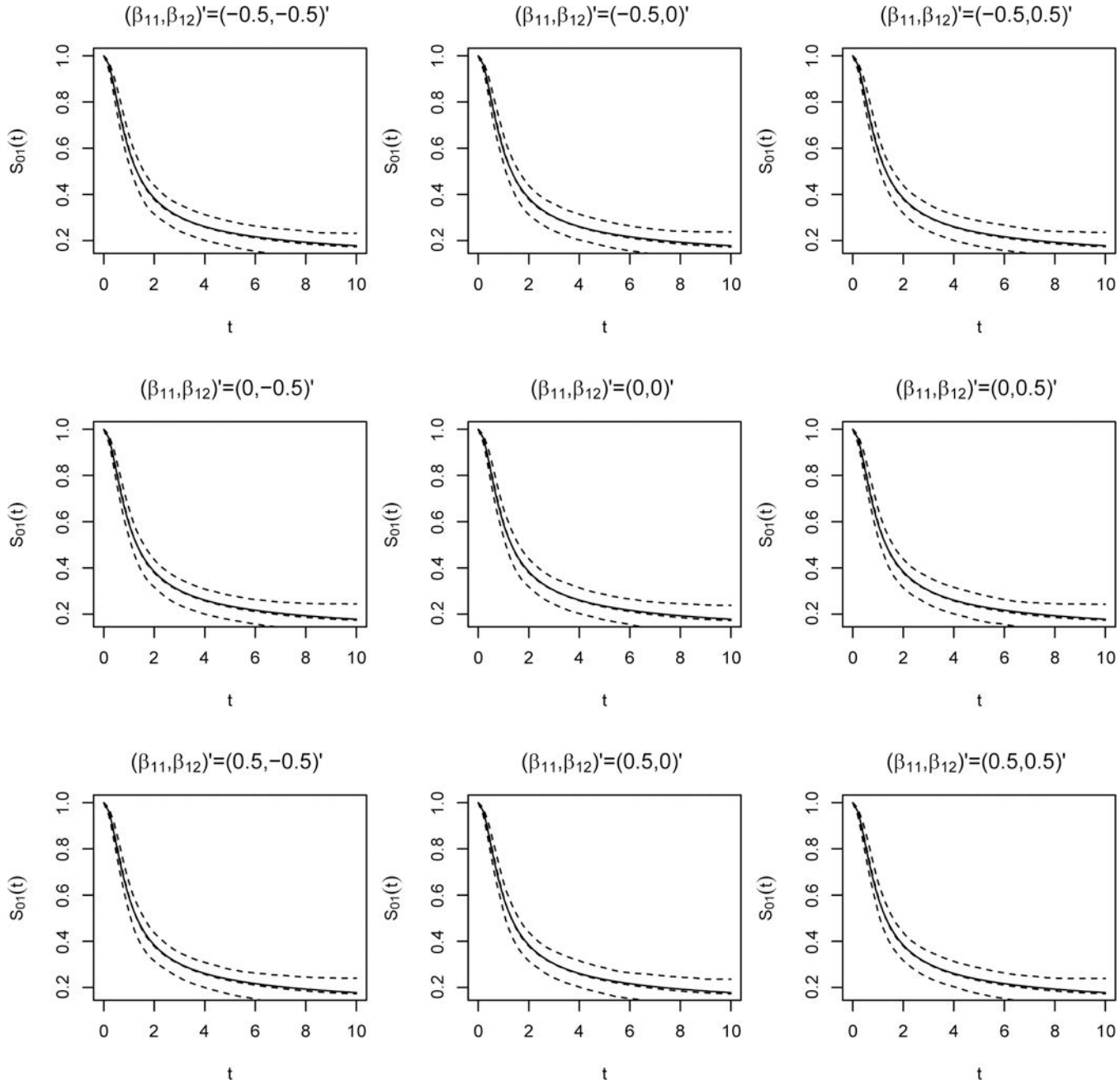

Figure 1:

Simulation results summarizing the estimates of the baseline survival function for failure time 1 obtained by the proposed approach, when $\nu = 1$, $m = 5$, and degree is three. The solid line provides the true value, dashed line represents the average estimated value, and the dotted lines indicate the 2.5% and 97.5% quantiles, of the point-wise estimates.

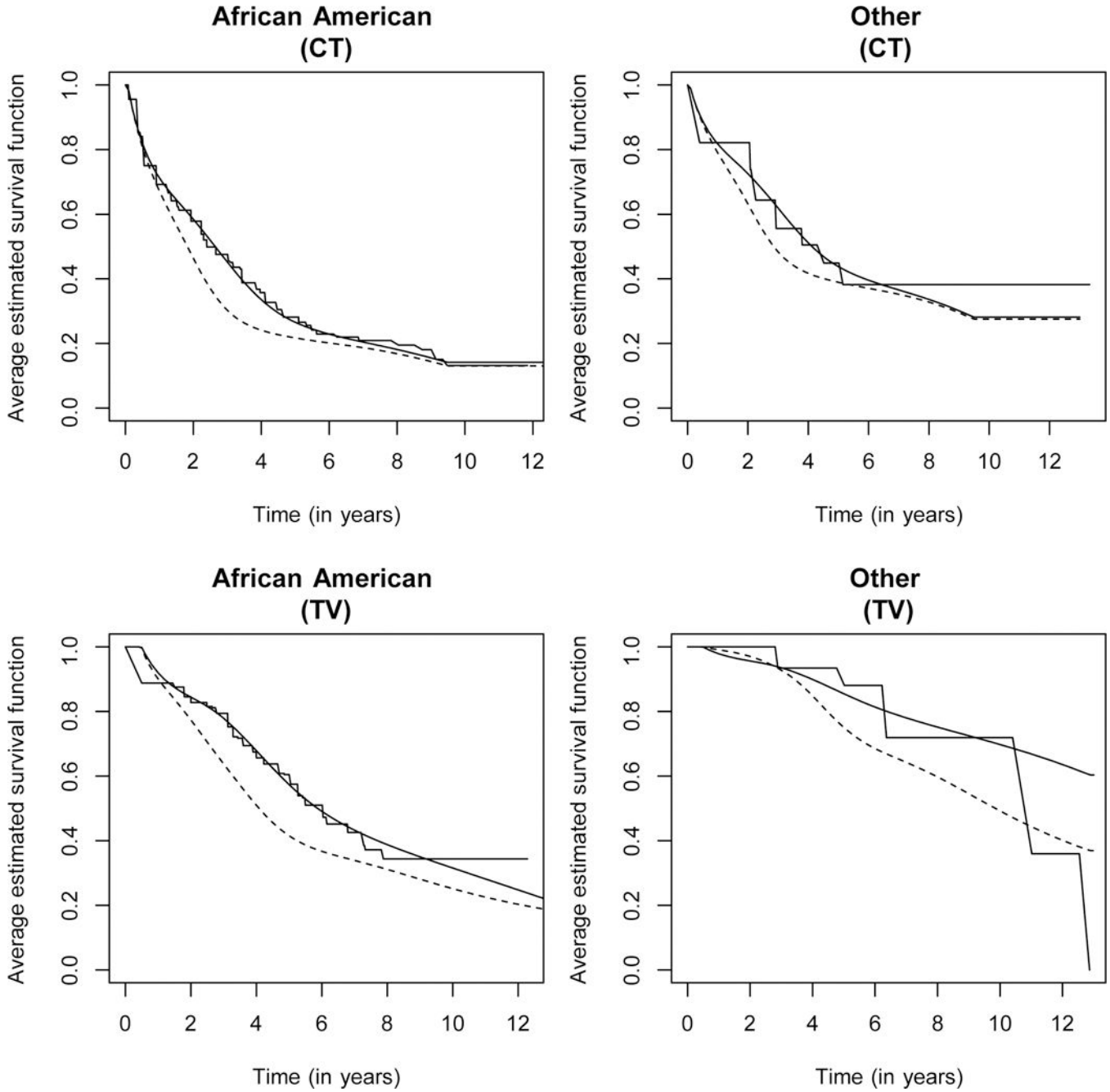

Figure 2:

Average estimated survival functions for CT and TV (stratified by race) obtained using the proposed method (solid smooth curves), the univariate/marginal method (dashed lines), and the Turnbull estimator (step functions).

Table 1:

Simulation results summarizing the estimates of the regression coefficients obtained from the proposed, univariate, and marginal methods, when $\nu = 1$, m = 5, and degree is three. This summary include the average of the 500 point estimates minus the true value (Bias), the sample standard deviation of the 500 point estimates (SD), the average of the estimated standard errors (ESE), the median of the estimated standard errors (mdSE) for the marginal approach only, and empirical coverage probabilities associated with 95% confidence intervals for the regression coefficients (CP95).

	Bivariate EM				Univariate method				Marginal method				
Parameter	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**mdSE**	**CP95**
$\beta_{11} = -0.5$	−0.04	0.18	0.18	0.94	−0.14	0.22	0.21	0.87	−0.14	0.22	0.87	0.23	1.00
$\beta_{12} = -0.5$	−0.05	0.19	0.18	0.93	−0.14	0.23	0.22	0.90	−0.14	0.23	0.39	0.22	1.00
$\beta_{21} = -0.5$	−0.04	0.17	0.18	0.96	−0.13	0.22	0.22	0.91	−0.13	0.22	0.34	0.23	0.99
$\beta_{22} = -0.5$	−0.03	0.18	0.18	0.94	−0.12	0.22	0.22	0.92	−0.12	0.22	0.24	0.23	0.94
$\beta_{11} = -0.5$	−0.04	0.18	0.18	0.95	−0.14	0.22	0.23	0.91	−0.14	0.22	0.46	0.24	1.00
$\beta_{12} = 0.0$	−0.01	0.19	0.18	0.95	−0.01	0.23	0.22	0.94	−0.01	0.23	0.23	0.22	0.96
$\beta_{21} = -0.5$	−0.04	0.17	0.18	0.96	−0.14	0.21	0.22	0.92	−0.14	0.21	0.38	0.24	1.00
$\beta_{22} = 0.0$	0.01	0.17	0.18	0.96	0.01	0.21	0.22	0.96	0.01	0.21	0.23	0.22	0.97
$\beta_{11} = -0.5$	−0.04	0.18	0.18	0.95	−0.13	0.22	0.21	0.90	−0.13	0.22	0.32	0.23	0.99
$\beta_{12} = 0.5$	0.02	0.19	0.18	0.95	0.10	0.23	0.22	0.92	0.10	0.23	0.23	0.22	0.93
$\beta_{21} = -0.5$	−0.03	0.18	0.18	0.96	−0.13	0.22	0.22	0.90	−0.13	0.22	0.31	0.23	0.98
$\beta_{22} = 0.5$	0.04	0.18	0.18	0.97	0.14	0.22	0.22	0.90	0.14	0.22	0.23	0.23	0.91
$\beta_{11} = 0.0$	−0.01	0.18	0.18	0.95	−0.01	0.21	0.22	0.96	−0.01	0.21	0.34	0.23	1.00
$\beta_{12} = -0.5$	−0.04	0.19	0.18	0.94	−0.13	0.22	0.22	0.90	−0.13	0.22	0.23	0.22	0.93
$\beta_{21} = 0.0$	−0.01	0.17	0.18	0.95	−0.01	0.21	0.22	0.95	−0.01	0.21	0.32	0.23	1.00
$\beta_{22} = -0.5$	−0.03	0.18	0.18	0.96	−0.11	0.21	0.22	0.93	−0.11	0.21	0.24	0.22	0.97
$\beta_{11} = 0.0$	−0.01	0.18	0.18	0.95	−0.01	0.21	0.22	0.96	−0.01	0.21	0.34	0.24	1.00
$\beta_{12} = 0.0$	−0.01	0.19	0.18	0.93	−0.01	0.23	0.22	0.93	−0.01	0.23	0.23	0.22	0.93
$\beta_{21} = 0.0$	0.00	0.17	0.18	0.95	−0.01	0.21	0.22	0.95	−0.01	0.21	0.33	0.24	1.00
$\beta_{22} = 0.0$	0.01	0.17	0.18	0.96	0.01	0.21	0.22	0.96	0.01	0.21	0.23	0.22	0.96
$\beta_{11} = 0.0$	−0.01	0.18	0.18	0.96	−0.01	0.21	0.22	0.95	−0.01	0.21	4.76	0.23	1.00
$\beta_{12} = 0.5$	0.02	0.18	0.18	0.95	0.10	0.22	0.22	0.94	0.10	0.22	1.67	0.22	1.00
$\beta_{21} = 0.0$	−0.01	0.17	0.18	0.97	−0.01	0.21	0.22	0.97	−0.01	0.21	0.40	0.23	1.00
$\beta_{22} = 0.5$	0.04	0.17	0.18	0.96	0.12	0.21	0.22	0.91	0.12	0.21	0.23	0.22	0.93
$\beta_{11} = 0.5$	0.02	0.18	0.18	0.96	0.07	0.19	0.22	0.97	0.07	0.19	0.54	0.23	1.00
$\beta_{12} = -0.5$	−0.04	0.19	0.19	0.94	−0.10	0.21	0.21	0.93	−0.10	0.21	0.24	0.22	0.96
$\beta_{21} = 0.5$	0.03	0.18	0.18	0.96	0.08	0.20	0.22	0.96	0.08	0.20	0.30	0.23	1.00
$\beta_{22} = -0.5$	−0.03	0.18	0.19	0.96	−0.08	0.20	0.21	0.96	−0.08	0.20	0.22	0.22	0.96

	Bivariate EM				Univariate method				Marginal method				
Parameter	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**mdSE**	**CP95**
$\beta_{11} = 0.5$	0.02	0.18	0.18	0.97	0.08	0.20	0.22	0.95	0.08	0.20	0.38	0.24	1.00
$\beta_{12} = 0.0$	−0.01	0.18	0.18	0.95	−0.01	0.22	0.21	0.94	−0.01	0.22	0.22	0.21	0.94
$\beta_{21} = 0.5$	0.03	0.18	0.18	0.95	0.09	0.20	0.22	0.96	0.09	0.20	0.38	0.24	1.00
$\beta_{22} = 0.0$	0.01	0.17	0.18	0.97	0.01	0.20	0.21	0.97	0.01	0.20	0.22	0.22	0.98
$\beta_{11} = 0.5$	0.03	0.18	0.18	0.95	0.08	0.20	0.21	0.95	0.08	0.20	0.30	0.23	1.00
$\beta_{12} = 0.5$	0.02	0.18	0.19	0.96	0.07	0.21	0.21	0.94	0.07	0.21	0.23	0.21	0.96
$\beta_{21} = 0.5$	0.02	0.18	0.18	0.95	0.07	0.21	0.22	0.95	0.07	0.21	0.28	0.23	0.99
$\beta_{22} = 0.5$	0.03	0.18	0.19	0.95	0.09	0.21	0.21	0.94	0.09	0.21	0.22	0.22	0.94

Table 2:

Simulation results summarizing the estimates of ν obtained from the proposed method, across all considered values of ν, when $m = 5$, and degree is three. This summary include the average of the 500 point estimates minus the true value (Bias), the sample standard deviation of the 500 point estimates (SD), the average of the estimated standard errors (ESE), and empirical coverage probabilities associated with 95% confidence intervals (CP95).

	$\nu = 0.25$				$\nu = 1$				$\nu = 4$			
Configuration	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**CP95**
$\beta_{11} = \beta_{21} = -0.5$ $\beta_{12} = \beta_{22} = -0.5$	−0.02	0.03	0.03	0.91	−0.07	0.15	0.15	0.94	0.22	0.22	2.09	0.96
$\beta_{11} = \beta_{21} = -0.5$ $\beta_{12} = \beta_{22} = 0.0$	−0.02	0.03	0.03	0.91	−0.07	0.16	0.15	0.94	0.09	1.77	1.71	0.94
$\beta_{11} = \beta_{21} = -0.5$ $\beta_{12} = \beta_{22} = 0.5$	−0.02	0.03	0.03	0.92	−0.07	0.16	0.15	0.95	0.16	1.93	1.79	0.95
$\beta_{11} = \beta_{21} = 0.0$ $\beta_{12} = \beta_{22} = -0.5$	−0.02	0.03	0.03	0.92	−0.07	0.15	0.15	0.94	0.30	4.73	3.24	0.98
$\beta_{11} = \beta_{21} = 0.0$ $\beta_{12} = \beta_{22} = 0.0$	−0.02	0.03	0.03	0.92	−0.07	0.16	0.15	0.93	0.02	1.94	1.69	0.96
$\beta_{11} = \beta_{21} = 0.0$ $\beta_{12} = \beta_{22} = 0.5$	−0.02	0.03	0.03	0.93	−0.07	0.16	0.15	0.93	0.22	4.90	3.07	0.98
$\beta_{11} = \beta_{21} = 0.5$ $\beta_{12} = \beta_{22} = -0.5$	−0.02	0.03	0.03	0.94	−0.06	0.15	0.15	0.95	0.33	4.89	3.48	0.98
$\beta_{11} = \beta_{21} = 0.5$ $\beta_{12} = \beta_{22} = 0.0$	−0.02	0.03	0.03	0.94	−0.06	0.15	0.15	0.95	0.36	5.41	3.67	0.99
$\beta_{11} = \beta_{21} = 0.5$ $\beta_{12} = \beta_{22} = 0.5$	−0.02	0.03	0.03	0.94	−0.05	0.16	0.15	0.94	0.30	4.90	3.24	0.98

Table 3:

Simulation results summarizing the estimates of the regression coefficients obtained from the proposed method, under the settings in the robustness study. This summary include the average of the 500 point estimates minus the true value (Bias), the sample standard deviation of the 500 point estimates (SD), the average of the estimated standard errors (ESE), and empirical coverage probabilities associated with 95% confidence intervals for the regression coefficients (CP95).

	frailty model 1				frailty model 2				frailty model 3			
Parameter	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**CP95**
$\beta_{11} = -0.5$	−0.02	0.17	0.18	0.96	−0.05	0.17	0.18	0.96	−0.02	0.14	0.14	0.95
$\beta_{12} = -0.5$	−0.01	0.18	0.18	0.95	−0.03	0.18	0.18	0.95	−0.03	0.14	0.14	0.95
$\beta_{21} = -0.5$	−0.02	0.18	0.18	0.96	−0.05	0.18	0.18	0.94	−0.03	0.14	0.14	0.95
$\beta_{22} = -0.5$	0.00	0.18	0.18	0.95	−0.04	0.19	0.18	0.95	−0.02	0.14	0.14	0.95
$\beta_{11} = -0.5$	−0.01	0.17	0.18	0.95	−0.05	0.17	0.18	0.96	−0.02	0.14	0.14	0.96
$\beta_{12} = 0.0$	0.01	0.17	0.18	0.97	0.02	0.18	0.18	0.95	−0.01	0.13	0.14	0.95
$\beta_{21} = -0.5$	−0.02	0.18	0.18	0.94	−0.05	0.19	0.18	0.95	−0.03	0.14	0.14	0.93
$\beta_{22} = 0.0$	0.01	0.17	0.17	0.96	0.01	0.18	0.18	0.97	0.00	0.13	0.14	0.96
$\beta_{11} = -0.5$	−0.01	0.17	0.18	0.96	−0.05	0.18	0.18	0.94	−0.02	0.13	0.14	0.97
$\beta_{12} = 0.5$	0.02	0.18	0.18	0.96	0.07	0.18	0.18	0.94	0.02	0.14	0.14	0.95
$\beta_{21} = -0.5$	−0.02	0.19	0.18	0.94	−0.04	0.18	0.18	0.96	−0.02	0.15	0.14	0.94
$\beta_{22} = 0.5$	0.02	0.17	0.18	0.95	0.05	0.18	0.18	0.94	0.02	0.14	0.14	0.96
$\beta_{11} = 0.0$	−0.01	0.16	0.17	0.96	0.00	0.17	0.18	0.97	0.01	0.13	0.14	0.96
$\beta_{12} = -0.5$	0.00	0.18	0.18	0.95	−0.03	0.18	0.19	0.95	−0.03	0.14	0.14	0.95
$\beta_{21} = 0.0$	−0.01	0.18	0.17	0.94	0.00	0.18	0.18	0.95	0.00	0.14	0.14	0.95
$\beta_{22} = -0.5$	0.00	0.18	0.18	0.94	−0.03	0.19	0.19	0.95	−0.02	0.13	0.14	0.98
$\beta_{11} = 0.0$	0.00	0.17	0.17	0.95	0.00	0.18	0.18	0.96	0.00	0.13	0.14	0.95
$\beta_{12} = 0.0$	0.01	0.17	0.17	0.95	0.01	0.18	0.18	0.96	0.00	0.14	0.14	0.95
$\beta_{21} = 0.0$	−0.01	0.17	0.17	0.95	0.01	0.19	0.18	0.96	0.00	0.14	0.14	0.95
$\beta_{22} = 0.0$	0.01	0.17	0.17	0.95	0.01	0.18	0.18	0.95	0.01	0.14	0.14	0.95
$\beta_{11} = 0.0$	0.00	0.16	0.17	0.95	0.00	0.18	0.18	0.95	0.01	0.13	0.14	0.96
$\beta_{12} = 0.5$	0.02	0.17	0.18	0.96	0.06	0.19	0.19	0.94	0.02	0.14	0.14	0.96
$\beta_{21} = 0.0$	−0.01	0.18	0.17	0.95	0.01	0.19	0.18	0.95	0.00	0.14	0.14	0.94
$\beta_{22} = 0.5$	0.01	0.17	0.18	0.94	0.05	0.19	0.19	0.94	0.03	0.14	0.14	0.94
$\beta_{11} = 0.5$	0.00	0.17	0.17	0.96	0.04	0.18	0.19	0.95	0.03	0.15	0.14	0.94
$\beta_{12} = -0.5$	0.00	0.17	0.18	0.96	−0.02	0.19	0.19	0.95	−0.03	0.14	0.15	0.95
$\beta_{21} = 0.5$	0.00	0.17	0.17	0.95	0.04	0.19	0.19	0.94	0.03	0.15	0.14	0.95
$\beta_{22} = -0.5$	0.00	0.18	0.18	0.95	−0.03	0.19	0.19	0.95	−0.03	0.14	0.15	0.96

	frailty model 1				frailty model 2				frailty model 3			
Parameter	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**CP95**	**Bias**	**SD**	**ESE**	**CP95**
$\beta_{11} = 0.5$	0.00	0.17	0.17	0.96	0.04	0.18	0.19	0.96	0.03	0.15	0.14	0.94
$\beta_{12} = 0.0$	0.01	0.17	0.17	0.96	0.02	0.18	0.19	0.96	0.00	0.14	0.14	0.96
$\beta_{21} = 0.5$	−0.01	0.17	0.17	0.95	0.04	0.19	0.19	0.94	0.03	0.14	0.14	0.95
$\beta_{22} = 0.0$	0.00	0.17	0.17	0.94	0.01	0.18	0.19	0.97	0.00	0.14	0.14	0.96
$\beta_{11} = 0.5$	0.00	0.17	0.17	0.96	0.05	0.19	0.19	0.94	0.03	0.14	0.14	0.94
$\beta_{12} = 0.5$	0.01	0.17	0.18	0.96	0.06	0.19	0.19	0.94	0.02	0.14	0.15	0.96
$\beta_{21} = 0.5$	−0.01	0.18	0.17	0.95	0.04	0.19	0.19	0.95	0.03	0.15	0.14	0.93
$\beta_{22} = 0.5$	0.02	0.17	0.18	0.95	0.04	0.19	0.19	0.94	0.03	0.14	0.15	0.95

Table 4:

STI data analysis: Estimated regression coefficients, estimated standard errors (ESE), and p-values obtained by the proposed, univariate, and the marginal methods.

		Bivariate EM			Univariate method			Marginal method		
	Covariate	**Estimate**	**ESE**	**p-value**	**Estimate**	**ESE**	**p-value**	**Estimate**	**ESE**	**p-value**
CT	No.of Partners (x_1)	0.1020	0.0322	0.0015	0.0982	0.0317	0.0019	0.0982	0.0532	0.0643
	Age at first coitus (x_2)	0.1587	0.0686	0.0209	0.1646	0.0991	0.0969	0.1646	0.2319	0.4777
	Race (x_3)	0.7043	0.2757	0.0108	0.6917	0.2996	0.0209	0.6917	0.4739	0.1443
TV	No.of Partners (x_1)	0.0483	0.0259	0.0629	0.0689	0.0508	0.1738	0.0689	0.0499	0.1676
	Age at first coitus (x_2)	−0.0299	0.0703	0.6672	−0.0067	0.0458	0.8808	−0.0067	0.1076	0.9522
	Race (x_3)	1.4605	0.4879	0.0028	2.7947	0.6705	<0.0001	2.7947	0.9157	0.0023