Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Analysis of Multivariate Interval Censoring by Diabetic Retinopathy Study

Man-Hua Chen , Li-Ching Chen , Kuen-Hung Lin & Xingwei Tong

To cite this article: Man-Hua Chen , Li-Ching Chen , Kuen-Hung Lin & Xingwei Tong (2014) Analysis of Multivariate Interval Censoring by Diabetic Retinopathy Study, Communications in Statistics - Simulation and Computation, 43:7, 1825-1835, DOI: 10.1080/03610918.2012.745557

To link to this article: https://doi.org/10.1080/03610918.2012.745557

Accepted author version posted online: 01 Oct 2013.
Published online: 01 Oct 2013.

Communications in Statistics—Simulation and Computation®, 43: 1825–1835, 2014

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2012.745557

Taylor & Francis
Taylor & Francis Group

Analysis of Multivariate Interval Censoring by Diabetic Retinopathy Study

MAN-HUA CHEN,[1] LI-CHING CHEN,[1] KUEN-HUNG LIN,[1] AND XINGWEI TONG[2]

[1]Department of Statistics, Tamkang University, Tamsui, Taiwan
[2]School of Mathematical Sciences, Beijing Normal University, Beijing, China

Multivariate failure time data are commonly encountered in biomedical research since each study subject may experience multiple events or because there exists clustering of subjects such that failure times within the same cluster are correlated. In this article, we use the frailty approach to catch the related survival variables and assume each event is a discrete analog as an interval of clinical examinations periodically. For estimation, an Expectation–Maximization (EM) algorithm is developed and is applied to the diabetic retinopathy study (DRS).

Keywords EM algorithm; Frailty models; Interval-censoring; Multivariate failure time data.

Mathematics Subject Classification 62N01.

1. Introduction

This article discusses the fitting of the frailty model to multivariate interval-censored data. One field in which interval-censored failure time data frequently occur is medical follow-up studies, and in these cases, each study subject is commonly examined or observed periodically. In this situation, an individual due to the pre-scheduled observations for a clinically observable change in disease or health status may miss some observations and return with a changed status. Accordingly, we only know that the true event time is greater than the last observation time at which the change has been observed not to occur, thus giving an interval that contains the real time of occurrence of the change. Goggins and Finkelstein (2000) presented a set of bivariate interval-censored data arising from an AIDS clinical trial on HIV-infected individuals. Kim and Xue (2002) discussed an ongoing clinical trial involving subjects with systemic lupus erythematosus. Chen et al. (2007) and Tong et al. (2008) developed the marginal model approach for multivariate interval-censored failure time data using the proportional odds model and the additive hazards model, respectively. Wang et al. (2008) discussed efficient estimation for bivariate current status data. Also Komarek and Lesaffre (2007) gave a Bayesian approach for correlated interval-censored data.

Received February 1, 2012; Accepted October 29, 2012
Address correspondence to Dr. Man-Hua Chen, Department of Statistics, Tamkang University, Tamsui, Taiwan; E-mail: mchen@mail.tku.edu.tw

Among these, two types of models have been proposed: frailty models and marginal models. Frailty model approach specifies the within-cluster correlation that allows for joint inference about the survival times within a cluster. Marginal models leave the correlation unspecified, but modify for the correlation by using sandwich-type estimators for the variance. Clayton and Cuzick (1985) extended the proportional hazards model and included a random effect representing heterogeneity of subjects. Oakes (1989) considered the frailty model for bivariate failure time data. Duchateau and Janssen (2008) pointed out that a drawback of the lognormal frailty distribution is that the Laplace transform does not take a simple form and hence the dependence imposed by the lognormal distribution is difficult to evaluate. In this article, our interest focused on survival prediction for interval-censored data under the Cox proportional hazards frailty model (Hougaard, 2000) and we use the frailty approach to catch the multi-events in an individual subject. Multi-events of the interval censoring from each individual share a common frailty random variable, which accounts for the within-individual correlation. Compared with the marginal model approach, one advantage of the frailty model approach is that it directly models the correlation of failure times. For estimation of the frailty model approach, an EM algorithm is developed (Klein, 1992).

We first present models and assumptions in Section 2. Section 3 discusses estimation of unknown parameters by maximizing the log-likelihood function of the pseudo-complete data. For this procedure, we use the EM algorithm. The EM algorithm iterates between an expectation and maximization step. The resulting estimates of regression parameters are consistent and asymptotically normally distributed. For the covariance matrix of the estimated parameters, a robust estimate is given that takes into account the correlation of the survival variables. In Section 4, some simulation results are presented and indicate that the presented inference approach works well for practical situations. We apply the approach to the diabetic retinopathy study (DRS) in Section 5. Section 6 contains some discussion.

2. Model and Assumptions

Consider a survival study that involves K possibly correlated failure times $(T_1, \ldots, T_K)$. Suppose that the T_k's can be observed only to belong to one of J different intervals given by or each study subject is observed only at J time points $0 = t_0 < t_1 < t_2 < \cdots < t_J < t_{J+1} = \infty$. For each subject, assume that there is a vector of covariates X_{ik} associated with the failure time T_{ik}, $i = 1, \ldots, n, k = 1, \ldots, K$. In the following, we assume that there exist K latent variables $b_{i1}, \ldots, b_{ik}$ for each subject and given b_{ik} and X_{ik}, the hazard function of T_{ik} has the form

$$\lambda_{ik}(t) = \lambda_{0k}(t) e^{X'_{ik}\beta + b_{ik}}, \tag{1}$$

where $i = 1, \ldots, n$, $\lambda_{0k}(t)$ denotes an unknown baseline hazard function, $k = 1, \ldots, K$, and β denotes vectors of regression parameters. That is, T_{ik} follows the proportional hazards frailty model. The simplest model is assuming the baseline survival functions for T_k's are the same as well as the covariates' effects on them. In contrast, allowing the baseline survival functions for T_k's is different and considering the effects of covariates on T_k's may be the same or different. The methodology given below still applies if covariate effects differ for T_k's as one can simply produce a common β through the introduction of extra type-specific covariates (Guo and Lin, 1994). For inference about β, we assume that only interval-censored data about the T_k's are available and they have the form

$\{(L_{ik}, R_{ik}],\ Z_i;\ i = 1, \ldots, n,\ k = 1, \ldots, K\}$. In the above, $(L_{ik}, R_{ik}]$ denotes the interval within which the kth failure of the ith subject is observed to occur. $(L_{ik}, R_{ik}]$ is general or case II interval-censored data (Huang and Wellner, 1997; Sun, 2006) for the kth failure of the ith subject T_{ik}. Here we use the convention that $L_{ik} = R_{ik}$ means that we have an exact observation on the kth failure time of the ith subject and $R_{ik} = t_{J+1} = \infty$ means that the observation on T_{ik} is right censored. In the following, we assume that $\{L_{ik}, R_{ik}\} \subseteq \{t_j\}$ and define $\alpha_{ikj} = 1$ if $(L_{ik}, R_{ik}]$ contains t_j and $\alpha_{ikj} = 0$ otherwise, $j = 1, \ldots, J+1$, $k = 1, \ldots, K$, $i = 1, \ldots, n$. Considering of the model which is denoted as the lognormal frailty model, we will assume that the latent effect $b = (b_1, \ldots, b_K)'$ follows a joint normal distribution with mean zero and covariance matrix Σ and given b, $T_1, \ldots, T_K$ are independent.

Under model (1), the probability T_k is observed to belong to the jth interval $(t_{j-1}, t_j]$ and is given by $P_{kj} = e^{-\Lambda_{0k}(t_{j-1})e^{X_k'\beta+b_k}} - e^{-\Lambda_{0k}(t_j)e^{X_k'\beta+b_k}}$, where $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(s)ds$, the unknown baseline cumulative hazard function. The Λ_{0k}'s subject to $0 \le \Lambda_{0k}(t_1) \le \cdots \le \Lambda_{0k}(t_J)$. In practice, it is convenient to eliminate the parameter range restriction by defining $\Delta_{km} = \Lambda_{0k}(t_j) - \Lambda_{0k}(t_{j-1})$ and $\gamma_{kj} = \log \Delta_{kj}$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Let $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kJ})'$ and $\gamma = (\gamma_1', \ldots, \gamma_K')'$. Then the likelihood contribution from a single subject has the form

$$L^*(\theta; O^*) = \int \prod_{k=1}^{K} \left\{ \alpha_{k1} + \sum_{j=1}^{J} \left(\alpha_{k(j+1)} - \alpha_{kj}\right) e^{-\sum_{a=1}^{j} e^{\gamma_{ka}} e^{X_k'\beta+b_k}} \right\} f(b; \Sigma)db, \tag{2}$$

where $\theta = (\beta, \Sigma, \gamma)$ notates all unknown parameters, $O^* = (\alpha_{kj}, X_k)$ the observed data, and $f(b; \Sigma)$ the density function of the normal distribution with mean zero and covariance Σ. From (2), one can easily show that the conditional density function of b given O^* has the form

$$f(b|O^*, \theta) = \frac{1}{L^*(\theta; O^*)} \prod_{k=1}^{K} \left\{ \alpha_{k1} + \sum_{j=1}^{J} \left(\alpha_{k(j+1)} - \alpha_{kj}\right) e^{-\sum_{a=1}^{j} e^{\gamma_{ka}} e^{X_k'\beta+b_k}} \right\} f(b; \Sigma). \tag{3}$$

In the next section, we discuss the estimation of regression parameters β along with other parameters.

3. Parameter Estimation

Suppose that the observed data O_i, $i = 1, \ldots, n$ are n iid copies of O^*. Then, we rewrite (2) into the full likelihood function as $L(\theta; O) = \prod_{i=1}^{n} L^*(\theta; O_i)$. To estimate θ, one needs to maximize $L(\theta; O)$. It is apparent that there is no closed form and there exists unobservable latent variables b's. It is a natural approach to apply the EM algorithm and treat b's as missing values.

3.1. *E-Step*

To describe the E-step, note that the pseudo-complete data consist of two parts: the observed data O_i and the missing data $b_i = \{b_{ik}\}_{k=1}^{K}$, $i = 1, \ldots, n$. The E-step asks to write out the likelihood function of the pseudo-complete data and then to compute the expectation of the resulting log-likelihood with respect to the conditional density function of b given O. It is easy to see that the log-likelihood function of the pseudo-complete data can be written as

$l(\theta; O, b) = \sum_{i=1}^{n} l_i(\theta; O_i, b_i)$, where

$$l_i(\theta; O_i, b_i) = \log f(b_i; \Sigma) + \sum_{k=1}^{K} \log \left\{ \alpha_{ik1} + \sum_{j=1}^{J} \big(\alpha_{ik(j+1)} - \alpha_{ikj}\big) e^{-\sum_{a=1}^{j} e^{\gamma_{ka}} e^{X'_{ik}\beta + b_{ik}}} \right\}.$$

It follows that the conditional expectation of the log-likelihood has the form

$$l(\theta; O) = \sum_{i=1}^{n} E[l_i(\theta; O_i, b_i)] = \sum_{i=1}^{n} \int l_i(\theta; O_i, b_i) f(b_i | O_i, \theta) db_i$$

with θ set to be $\theta^{(m)}$ obtained at the mth iteration. It is obvious that to compute the conditional expectation above, one requires a numerical algorithm to assess the general integral

$$E\big(h(b_i)|O_i, \theta^{(m)}\big) = \int h(b_i) f\big(b_i|O_i, \theta^{(m)}\big) db_i$$

for any function $h(b_i)$ of b_i.

For the determination of $E(h(b_i)|O_i, \theta^{(m)})$, we have

$$E\big(h(b_i)|O_i, \theta^{(m)}\big) = \frac{E[\psi(b_i; \theta^{(m)}, O_i) h(b_i)]}{E\psi(b_i; \theta^{(m)}, O_i)},$$

where $\psi(b_i; \theta^{(m)}, O_i) = \prod_{k=1}^{K} \{\alpha_{ik1} + \sum_{j=1}^{J} (\alpha_{ik(j+1)} - \alpha_{ikj}) e^{-\sum_{a=1}^{j} e^{\gamma_{ka}} e^{X'_{ik}\beta + b_{ik}}}\}$. This suggests that for sufficiently large L, the expectation $E(h(b_i)|O_i, \theta^{(m)})$ can be approximated by

$$E(h(b_i)|O_i, \theta^{(m)}) \simeq \widehat{E}(h(b_i)) = \frac{\sum_{l=1}^{L} \psi(b_{il}; \theta^{(m)}, O_i) h(b_{il})}{\sum_{l=1}^{L} \psi(b_{il}; \theta^{(m)}, O_i)}, \tag{4}$$

where $\{b_{il} = (b_{il1}, \ldots, b_{ilK})\}_{l=1}^{L}, i = 1, \ldots, n$ are iid samples from the K-dimensional normal distribution with mean zero and covariance matrix, $\Sigma^{(m)}$.

3.2. *M-Step*

First, we maximize the conditional expectation $l(\theta; O, b)$ by replacing all expectations involving functions $h(b)$ by their approximation $\widehat{E}(h(b))$ given in (4) to determine the updated estimate $\theta^{(m+1)}$. By taking derivatives of $l(\theta; O)$ with respect to Σ, one can easily obtain the updated estimator of Σ as

$$\Sigma^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{E}\big(b_i b_i' | O_i, \hat{\theta}^{(m)}\big). \tag{5}$$

For the maximum likelihood estimator of parameters β and γ, we have

$$U_\beta(\beta, \gamma) = \sum_{i=1}^{n} \sum_{k=1}^{K} E\big[\psi_{ik}^{(1)}(b_i; \beta, \gamma_k) | O_i, \hat{\theta}^{(m)}\big],$$

$$U_{\gamma_{kj}}(\beta, \gamma) = \sum_{i=1}^{n} E\big[\psi_{ikj}^{(2)}(b_i; \beta, \gamma_k) | O_i, \hat{\theta}^{(m)}\big],$$

where $\psi_{ik}^{(1)}(b_i;\beta,\gamma_k) = W_{ik}^{-1}V_{\beta,ik}$, $\psi_{ikj}^{(2)}(b_i;\beta,\gamma_k) = W_{ik}^{-1}V_{\gamma,ikj}$,

$$W_{ik} = \alpha_{ik1} + \sum_{j=1}^{J}\left(\alpha_{ik(j+1)} - \alpha_{ikj}\right)\left[e^{-e^{(X'_{ik}\beta+b_{ik})}\sum_{a=1}^{j}e^{\gamma_{ka}}}\right],$$

$$V_{\beta,ik} = \frac{\partial W_{ik}}{\partial\beta} = -X_{ik}\sum_{j=1}^{J}\left(\alpha_{ik(j+1)} - \alpha_{ikj}\right)\left[e^{-e^{(X'_{ik}\beta+b_{ik})}\sum_{a=1}^{j}e^{\gamma_{ka}}}\right]\left[e^{(X'_{ik}\beta+b_{ik})}\sum_{a=1}^{j}e^{\gamma_{ka}}\right],$$

$$V_{\gamma,ikj} = \frac{\partial W_{ik}}{\partial\gamma_{kj}} = -\sum_{s=j}^{J}\left(\alpha_{ik(s+1)} - \alpha_{iks}\right)e^{-e^{(X'_{ik}\beta+b_{ik})}\sum_{a=1}^{s}e^{\gamma_{ka}}}e^{X'_{ik}\beta+b_{ik}+\gamma_{kj}}.$$

Applying the approximation $\widehat{E}$ given in (4) and noticing that the denominator in (4) is a constant, we obtain the working score functions as $\widehat{U}_\beta(\beta,\gamma) = \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\sum_{l=1}^{L}\psi_{ik}^{(1)}(b_l;\beta,\gamma_k)\psi(b_l;\theta^{(m)},O_i)}{\sum_{l=1}^{L}\psi(b_l;\theta^{(m)},O_i)}$, $\widehat{U}_{\gamma_{kj}}(\beta,\gamma) = \sum_{i=1}^{n}\frac{\sum_{l=1}^{L}\psi_{ikj}^{(2)}(b_l;\beta,\gamma_k)\psi(b_l;\theta^{(m)},O_i)}{\sum_{l=1}^{L}\psi(b_l;\theta^{(m)},O_i)}$. Then for estimation of β and γ, one can solve the equation

$$\widehat{U}(\beta,\gamma) = (\widehat{U}_\beta(\beta,\gamma)', \widehat{U}_{\gamma_1}(\beta,\gamma_1)', \ldots, \widehat{U}_{\gamma_K}(\beta,\gamma_K)')' = 0, \tag{6}$$

where $\widehat{U}_{\gamma_k}(\beta,\gamma) = (\widehat{U}_{\gamma_{k1}}(\beta,\gamma), \ldots, \widehat{U}_{\gamma_{kJ}}(\beta,\gamma))'$.

3.3. *Computational Algorithm*

It is not easy to handle $p + K \times J$ equations in (6) simultaneously. Therefore, we suggest the following procedure for the $(m+1)$th iteration.

Step 1. Determine the updated estimate $\widehat{\Sigma}^{(m+1)}$ of Σ given in (5).
Step 2. Determine the updated estimate $\widehat{\beta}^{(m+1)}$ of β by solving $\widehat{U}_\beta(\beta,\gamma^{(m)}) = 0$.
Step 3. For each k, determine the updated estimate $\widehat{\gamma}_k^{(m+1)}$ of γ_k by solving $\widehat{U}_{\gamma_k}(\widehat{\beta}^{(m+1)},\gamma_k) = 0$, $k = 1,\ldots,K$.
Step 4. Repeat steps 1–3 until convergence.

For the covariance estimation of the estimator of $\widehat{\theta}$, one can use the inverse of the observed information matrix $I(\widehat{\theta})$, which is given in the Appendix. Also the distribution of $\widehat{\beta}$ or $\widehat{\theta}$ can be approximated by the normal distribution for large samples.

4. Simulation Study

Simulation studies were conducted to assess the finite sample performance of the proposed maximum likelihood estimators. In the study, we considered the situation where there exist $K = 2$ correlated failure times T_1 and T_2, and a scale covariate $X_1 = X_2 = X$ taking value 0 or 1 with probability 0.5 . The paired frailities were generated from the bivariate normal distribution with mean zero and covariance matrix Σ. In the above, their correlation ρ measures the dependence between T_1 and T_2 ; T_1 and T_2 with the hazard functions $0.1\,t_1\,e^{X'\beta+b_1}$ and $0.2\,t_2\,e^{X'\beta+b_2}$, respectively. In the study, we took each time point $t_j = j/J \times \tau$ (equally spaced partition) with $J = 5$ ($J \simeq n^{1/3}$) and $\tau = t_J$. More comments on J and the t_j's are given in Sun (2006). At each time point, a subject is observed with probability 1—censored rate and independent of observations at the other time points. Thus, L_{ik} and R_{ik} are the actual observation times that are immediately before and after the true

Table 1
Estimates of regression parameter β with $\rho = 0$

n	β	SD (b_1)	SD (b_2)	Censor rate	AVE	ESE	SSD	CP
100	(0,0)	0.4	0.4	0.2	(−.009, .002)	(.244, .221)	(.246, .231)	(.94, .95)
				0.5	(−.012, −.026)	(.270, .244)	(.282, .244)	(.94, .96)
			0.2	0.2	(.007, −.008)	(.245, .220)	(.251, .227)	(.96, .95)
				0.5	(.001, −.002)	(.271, .245)	(.272, .255)	(.95, .96)
	(0.5, 0.5)	0.4	0.4	0.2	(.472, .449)	(.247, .224)	(.253, .233)	(.94, .94)
				0.5	(.481, .462)	(.274, .249)	(.261, .265)	(.96, .93)
			0.2	0.2	(.465, .501)	(.247, .225)	(.243, .231)	(.95, .94)
				0.5	(.484, .510)	(.275, .251)	(.271, .259)	(.95, .95)
	(1, 0.5)	0.4	0.4	0.2	(.955, .480)	(.251, .227)	(.264, .223)	(.94, .94)
				0.5	(.958, .476)	(.281, .252)	(.321, .273)	(.93, .94)
			0.2	0.2	(.954, .488)	(.251, .227)	(.264, .233)	(.93, .95)
				0.5	(.953, .520)	(.280, .254)	(.293, .267)	(.93, .95)
200	(0, 0)	0.4	0.4	0.2	(−.014, .000)	(.171, .155)	(.177, .156)	(.95, .94)
				0.5	(.002, −.025)	(.189, .170)	(.192, .168)	(.95, .95)
			0.2	0.2	(.011, .011)	(.172, .154)	(.172, .157)	(.95, .94)
				0.5	(.003, −0.001)	(.190, .170)	(.187, .181)	(.94, .94)
	(0.5, 0.5)	0.4	0.4	0.2	(.481, .460)	(.173, .157)	(.175, .168)	(.94, .93)
				0.5	(.477, .445)	(.191, .173)	(.199, .173)	(.94, .94)
			0.2	0.2	(.467, .506)	(.173, .157)	(.173, .160)	(.95, .95)
				0.5	(.466, .484)	(.191, .174)	(.192, .194)	(.96, .93)
	(1, 0.5)	0.4	0.4	0.2	(.944, .464)	(.175, .159)	(.180, .158)	(.93, .94)
				0.5	(.936, .464)	(.195, .176)	(.187, .186)	(.95, .95)
			0.2	0.2	(.927, .489)	(.175, .159)	(.175, .162)	(.94, .95)
				0.5	(.920, .488)	(.194, .177)	(.191, .177)	(.93, .94)

failure time T_{ik}. The results given below are based on 500 replications with $L = 30$ for the approximation (4) and the sample size $n = 100$ or 200.

To evaluate $\hat{\beta}$, we considered a number of scenarios, and some of the obtained results are presented in Tables 1–3. In all tables, we included the averages of the estimates $\hat{\beta}$ (AVE), sample standard deviation of the 500 sample estimators (SSD), the average of the 500 estimated standard errors (ESE), and the coverage probability with confidence level 95% (CP). The true values of identical covariates β took 0, or 0.5, different covariates β took 1, and 0.5, the censor rate set to 0.2, or 0.5, and the standard deviations of b_1 and b_2 were set to be equal (0.4, 0.4) or different (0.4, 0.2). Table 1 studied the situation with $\rho = 0$, Table 2 assumed $\rho = 0.25$, and Table 3 assumed $\rho = 0.5$. One can see from these tables that the proposed estimate $\hat{\beta}$ seems to be unbiased and the variance estimate is reasonably reliable as it is close to the sample standard deviation. Also the results seem robust for different cases and especially, it is interesting to note that the variance estimates do not seem to change much from cases with $\rho = 0$ to cases with $\rho = 0.5$. The results also indicate that the ESEs and SSDs of $\hat{\beta}$ increased as the censored rate increased. This is due to the larger censored rate 0.5 corresponds to wide observed intervals for the survival times of interest and thus means less information about the survival times. Furthermore, as expected, both the bias and the estimated standard error decreased as the sample size increased. In the simulation study, we also considered other set-ups such as those with

Table 2
Estimates of regression parameter β with $\rho = 0.25$

n	β	SD(b_1)	SD(b_2)	Censor rate	AVE	ESE	SSD	CP
100	(0,0)	0.4	0.4	0.2	(−.006, .005)	(.245, .220)	(.265, ,230)	(.94, .94)
				0.5	(.002, −.002)	(.271, .244)	(.274, .252)	(.94, .95)
			0.2	0.2	(−.027, −.002)	(.245, .220)	(.252, .215)	(.95, .96)
				0.5	(.015, −.006)	(.272, .245)	(.286, .260)	(.95, .94)
	(0.5,0.5)	0.4	0.4	0.2	(.472, .466)	(.247, .224)	(.255, .236)	(.94, .94)
				0.5	(.464, .482)	(.274, .256)	(.283, .257)	(.95, .95)
			0.2	0.2	(.453, .489)	(.247, .225)	(.251, .224)	(.94, .95)
				0.5	(.479, .531)	(.274, .252)	(.274, .262)	(.95, .94)
	(1,0.5)	0.4	0.4	0.2	(.939, .483)	(.250, .226)	(.255, .232)	(.95, .95)
				0.5	(.963, .501)	(.281, .253)	(.282, .269)	(.94, .94)
			0.2	0.2	(.944, .502)	(.251, .227)	(.252, .228)	(.94, .96)
				0.5	(.935, .504)	(.278, .254)	(.274, .275)	(.94, .94)
200	(0,0)	0.4	0.4	0.2	(−.003, −.008)	(.171, .154)	(.181, 152)	(.94, .95)
				0.5	(−.012, .004)	(.189, .170)	(.193, .178)	(.95, .92)
			0.2	0.2	(.012,−.007)	(.172, 154)	(.183, .160)	(.93, .94)
				0.5	(.005, −.003)	(.190, .170)	(.183, .164)	(.96, .96)
	(0.5,0.5)	0.4	0.4	0.2	(.468, .471)	(.173, .157)	(.176, .154)	(.94, .96)
				0.5	(.472, .460)	(.192, .173)	(.201, .170)	(.94, .95)
			0.2	0.2	(.477, .499)	(.173, .157)	(.182, .157)	(.93, .96)
				0.5	(.471, .491)	(.191, .174)	(.197, .174)	(.94, .95)
	(1,0.5)	0.4	0.4	0.2	(.943, .469)	(.175, .159)	(.174, .160)	(.94, .95)
				0.5	(.937, .450)	(.195, .176)	(.193, .176)	(.94, .94)
			0.2	0.2	(.941, .494)	(.176, .159)	(.182, .159)	(.93, .94)
				0.5	(.949, .505)	(.195, .177)	(.198, .175)	(.95, .95)

larger L in (4), standard deviations of b_1 and b_2, and $J = 10$ ($J \simeq n^{1/2}$). Similar results were obtained.

A referee suggested comparing the proposed approach with a univariate approach. To see this, we applied this univariate approach to situations considered in Table 4 and presented the obtained results, including AVE, SEE, and SSE. It can be seen that as expected, the approach is less efficient than the estimation procedure developed here with higher correlation between T_1 and T_2.

5. Analysis of Diabetic Retinopathy Study

For illustration purposes, we applied our proposed method to the DRS (Huster et al., 1989). The study was conducted by the National Eye Institute to assess the effect of laser photocoagulation in delaying the onset of severe visual loss such as blindness in the patients with diabetic retinopathy, where 50% subsamples are the high-risk patients with $n = 197$. The data were categorized into 16 intervals, {(0, 6], (6, 10], (10, 14], . . . , (58, 66], (66, 83]} (in months). The main purpose of the study is to evaluate the outcome of laser photocoagulation in delaying the time to onset of blindness in patients with diabetic retinopathy. Defined covariates included in the analyses were type of diabetes ($x_1 = 0$ for juvenile diabetes and $x_1 = 1$ for adult diabetes) and presence/absence of treatment ($x_2 = 0$ if the patient received

Table 3
Estimates of regression parameter β with $\rho = 0.5$

n	β	SD(b_1)	SD(b_2)	Censor rate	AVE	ESE	SSD	CP
100	(0,0)	0.4	0.4	0.2	(.001, −.013)	(.244, .220)	(.224, .221)	(.96, .94)
				0.5	(.009, −.024)	(.271, .244)	(.275, .254)	(.94, .94)
			0.2	0.2	(−.004, .009)	(.245, .220)	(.264, .225)	(.92, .96)
				0.5	(−.001, −.002)	(.271, .245)	(.266, .249)	(.96, .95)
	(0.5,0.5)	0.4	0.4	0.2	(.474, .472)	(.247, .224)	(.250, .229)	(.95, .93)
				0.5	(.503, .470)	(.273, .250)	(.277, .268)	(.94, .93)
			0.2	0.2	(.480, .479)	(.247, .224)	(.251, .233)	(.95, .95)
				0.5	(.480, .509)	(.275, .251)	(.272, .267)	(.95, .94)
	(1,0.5)	0.4	0.4	0.2	(.925, .471)	(.250, .227)	(.236, .228)	(.95, .94)
				0.5	(.971, .479)	(.280, .252)	(.294, .256)	(.95, .95)
			0.2	0.2	(.947, .515)	(.251, .227)	(.244, .236)	(.95, .93)
				0.5	(.991, .499)	(.282, .253)	(.297, .250)	(.94, .96)
200	(0,0)	0.4	0.4	0.2	(.002, .006)	(.171, .154)	(.156, .156)	(.96, .95)
				0.5	(.002, −.004)	(.190, .170)	(.197, .179)	(.94. .95)
			0.2	0.2	(.008, .006)	(.172, .154)	(.177, .151)	(.94, .95)
				0.5	(−.004, −.008)	(.190, .170)	(.187, .172)	(.96, .95)
	(0.5,0.5)	0.4	0.4	0.2	(.470, .470)	(.173, .157)	(.178, .149)	(.94, .96)
				0.5	(.478, .473)	(.192, .174)	(.192, .184)	(.95, .95)
			0.2	0.2	(.455, .507)	(.173, .157)	(.173, .154)	(.95, .96)
				0.5	(.492, .500)	(.192, .175)	(.191, .169)	(.96, .95)
	(1,0.5)	0.4	0.4	0.2	(.927, .487)	(.175, .159)	(.171, .161)	(.94, .95)
				0.5	(.934, .474)	(.195, .176)	(.201, .171)	(.94, .96)
			0.2	0.2	(.920, .497)	(.175, .159)	(.174, .162)	(.94, .95)
				0.5	(.925, .496)	(.195, .177)	(.187, .184)	(.95, .95)

Table 4
Estimates of regression parameter $\beta = (1, 0.5)$ with univariate approach

SD(b_1)	SD(b_2)	ρ	n	Censor rate	AVE	ESE	SSD	CP
0.4	0.4	0	100	0.2	(.961, .462)	(.251, .227)	(.249, .232)	(.95, .94)
				0.5	(.940, .477)	(.280, .252)	(.280, .247)	(.93, .95)
			200	0.2	(.932, .460)	(.175, .159)	(.166, .149)	(.93, .95)
				0.5	(.942, .481)	(.195, .176)	(.188, .177)	(.96, .95)
		0.25	100	0.2	(.944, .450)	(.251, .227)	(.264, .221)	(.93, .96)
				0.5	(.908, .433)	(.278, .252)	(.253, .249)	(.93, .94)
			200	0.2	(.917, .445)	(.174, .159)	(.188, .172)	(.92, .92)
				0.5	(.887, .431)	(.193, .175)	(.194, .182)	(.90, .93)
		0.5	100	0.2	(.882, .467)	(.250, .227)	(.272, .254)	(.93, .93)
				0.5	(.889, .417)	(.277, .251)	(.320, .232)	(.89, .96)
			200	0.2	(.873, .424)	(.173, .158)	(.166, .136)	(.92, .93)
				0.5	(.905, .437)	(.195, .175)	(.208, .175)	(.89, .96)

no treatment and $x_2 = 1$ if the patient was treated with laser photocoagulation). Variable x_3 along with those from Huster et al. (1989) and Ross and Moore (1999) is interaction of two factors (x_1 and x_2). Among these, we considered the simplest and assumed that the baseline survival functions for T_1 and T_2 as well as the covariates' effects on them for the eye studies are the same.

In this article, we use the proposed method, the estimates of regression parameters are 0.364, −0.443, and −0.977. The p-values for type of diabetes (x_1), treatment (x_2), and interaction (x_3) are all smaller than 0.0001. The laser treatment appears to be effective; for juvenile diabetes, their risk of failure is reduced by 36% ($\exp(-0.45) = 0.64$) relative to the control group. The results also indicate that the laser treatment was more effective in the adult onset group than in those individuals with juvenile onset diabetes.

In this study, the rank for the non-right-censored survival times is missing, the midpoint (average of lowerbound and upperbound of the interval) is used for simplicity; some authors considered if it is possible to impute the exact survival times to each finite censored interval. The idea behind the imputation method is to change to a more familiar or simpler model. Our proposed method instead of dealing with prior information or data to suggest or verify a parametric model deals with interval censoring directly. Thus, we may avoid underestimation of the variability of point estimates (Sun, 2006).

6. Concluding Remarks

The method developed in the preceding sections applies to covariates measured at a single point in time. The extension to time-varying covariates is also possible as long as the covariates are ancillary or external. Chen and Tong (2010) proposed a maximum likelihood method with spline smoothing for varying coefficients. When a covariate interacts nonlinearly with another, one may extend our proposed method with varying coefficients to catch the phenomenon of the time-dependent covariates. Another question for the regression analysis of multivariate interval-censored data is how one can choose an appropriate model among all available models. The DRS that we discussed in Section 5 had been studies by several authors with different approaches. All these results point out that the estimated parameters are significant. There does not seem to exist an approach in the literature that can be used to choose or distinguish these different models and it is apparent that the development of such approach would be very useful.

Petroni and Wolfe (1994) considered that the survival time has a discrete distribution, although many methods of analysis assume that it has a continuous distribution. The discrete time scale can always be constructed from a continuous one by partitioning the time axis into disjoint intervals. In practice, the number of intervals, J, will depend on the amount of data available in order to maintain efficiency of the parameter estimates. It is easy to see that for larger J, more computational effort is needed, but a better approximation is obtained. Rossini and Tsiatis (1996) suggested the smallest integer of J above $n^{1/4}$, but for small n, one may want to choose larger J since the results may not be stable otherwise. Sun (2006) investigated the effect of the partition or selection of t_j's on the analysis of the lung tumor data and the results are quite stable through different J's.

Our proposed method provided an easy and simple way to deal with the dependence structure for more than two correlated failure times. We do not really estimate the parameters of the correlation; in practices, we may only focus on regression parameters. The performance of our proposed method is still good under there is no correlation between

failure times. In other words, one can use our proposed method even though the data may not exist the dependence structure or the correlation is weak.

Simulation experiments indicate that the estimates of regression parameters using the proposed EM approach are quite robust to the initial value of the standard deviations of frailties (SD($b1$), SD($b2$)), although a good initial value of frailties can certainly improve the convergence speed. The simulation experiences show that the performance of the estimation of frailties can be improved by increasing the sample size. In some cases, the strength of dependence between the failure time is of biological interest, and thus further efforts will be required in future work to provide robust inference of the standard deviations of frailties.

In addition to the cumulative hazard function, $\Lambda_{0k}(t)$ belongs to a linear function space, in some situations, it possibly prefers to use different spaces or a similar space in which the dimension J could be infinity. In this case, a new method needs to be developed for the estimation of covariate effects.

Acknowledgments

Dr. M.-H. Chen was supported in part by the National Science Council, Taiwan (Project No. NSC-97-2118-M-032-013). Dr. X. Tong was supported in part by the NSF China Zhongdian (Project No. 11131002) and NSFC (Project No. 10971015), and the Fundamental Research Funds for the Central Universities.

Appendix: Expression of the Observed Fisher Information Matrix

$I(\widehat{\theta})$ denotes the observed Fisher information matrix and has the form $-E(\frac{\partial^2 l(\theta;O,b)}{\partial\theta\theta'}|O,\widehat{\theta})$. Then $\frac{\partial^2 l(\theta;O)}{\partial\beta\partial\beta'} = \sum_{i=1}^{n}\sum_{k=1}^{K} E[W_{ik}^{-2}V_{\beta,ik}^2 - W_{ik}^{-1}V_{\beta\beta,ik}]$, $\frac{\partial^2 l(\theta;O)}{\partial\gamma_{kj}\partial\gamma_{kj}} = \sum_{i=1}^{n} E[W_{ik}^{-2}V_{\gamma,ikj}^2 - W_{ik}^{-1}V_{\gamma_j\gamma_j,ikj}]$, $\frac{\partial^2 l(\theta;O)}{\partial\gamma_{kj}\partial\beta} = \sum_{i=1}^{n} E[W_{ik}^{-2}V_{\gamma,ikj}V_{\beta,ik} - W_{ik}^{-1}V_{\gamma_j\beta,ikj}]$, and $\frac{\partial^2 l(\theta;O)}{\partial\gamma_{kj}\partial\gamma_{km}} = \sum_{i=1}^{n} E[W_{ik}^{-2}V_{\gamma,ikj}V_{\gamma,ikm} - W_{ik}^{-1}V_{\gamma_j\gamma_m,ikjm}]$, for $j < m$, where

$$
\begin{aligned}
V_{\beta\beta,ik} &= \sum_{j=1}^{J}\left(\alpha_{ik(j+1)} - \alpha_{ikj}\right)xx'\left(e^{(x'\beta+b_k)}\sum_{a=1}^{j}e^{\gamma_{ka}}\right)\\
&\quad\times\left(e^{(x'\beta+b_k)}\sum_{a=1}^{j}e^{\gamma_{ka}} - 1\right)\left(e^{-e^{(x'\beta+b_k)}\sum_{a=1}^{j}e^{\gamma_{ka}}}\right),\\
V_{\gamma_j\gamma_j,ikj} &= \sum_{s=j}^{J}\left(\alpha_{ik(s+1)} - \alpha_{iks}\right)\left(e^{(x'\beta+b_k)}e^{\gamma_{kj}} - 1\right)\\
&\quad\times\left(e^{-e^{(x'\beta+b_k)}\sum_{a=1}^{s}e^{\gamma_{ka}}}\right)\left(e^{(x'\beta+b_k)}e^{\gamma_{kj}}\right),\\
V_{\beta\gamma_j,ikj} &= \sum_{s=j}^{J}\left(\alpha_{ik(s+1)} - \alpha_{iks}\right)x\left(e^{(x'\beta+b_k)}e^{\gamma_{kj}}\right)\\
&\quad\times\left(e^{-e^{(x'\beta+b_k)}\sum_{a=1}^{s}e^{\gamma_{ka}}}\right)\left(e^{(x'\beta+b_k)}\sum_{a=1}^{s}e^{\gamma_{ka}} - 1\right),
\end{aligned}
$$

$$V_{\gamma_j\gamma_m,ikjm} = \sum_{s=j}^{J} \left(\alpha_{ik(s+1)} - \alpha_{iks}\right)\left(e^{(x'\beta+b_k)}e^{\gamma_{kj}}\right)$$

$$\times \left(e^{(x'\beta+b_k)}e^{\gamma_{km}}\right)\left(e^{-e^{(x'\beta+b_k)}\sum_{a=1}^{s} e^{\gamma_{ka}}}\right).$$

References

Chen, K., Tong, X. (2010). Varying coefficient transformation models with censored data. *Biometrika* 97:969–976.

Chen, M.-H., Tong, X., Sun, J. (2007). The proportional odds model for multivariate interval-censored failure time data. *Statistics in Medicine* 26:5147–5161.

Clayton, D., Cuzick, J. (1985). Multivariate generations of the proportional hazards model. *Journal of the Royal Statistical Society: Series A* 148:82–117.

Duchateau, L., Janssen, P. (2008). *The Frailty Model*. New York: Springer.

Goggins, W. B., Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* 56:940–943.

Guo, S. W., Lin, D. Y. (1994). Regression analysis of multivariate grouped survival data. *Biometrics* 50:632–639.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer.

Huang, J., Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In: Lin, D., Fleming, T., eds. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*. New York: Springer, pp. 123–169.

Huster, W. J., Brookmeyer, R., Self, S. G. (1989). Modeling paired survival data with covariates. *Biometrics* 45:145–156.

Kim, M. Y., Xue, X. (2002). The analysis of multivariate interval-censored survival data. *Statistics in Medicine* 21:3715–3726.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* 48:795–806.

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

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84:487–493.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics* 22:86–91.

Petroni, G. R., Wolfe, R. A. (1994). A two-sample test for stochastic ordering with interval-censored data. *Biometrics* 50:77–87.

Ross, E. A., Moore, D. (1999). Modeling clustered, discrete, or grouped time survival data with covariates. *Biometrics* 55:813–819.

Rossini, A. J., Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association* 91:713–721.

Sun, J. (2006). *The Statistical Analysis of Interval-censoring Failure Time Data*. New York: Springer.

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.

Turnbull, B. W. (1976). The empirical distribution function from arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B* 38:290–295.

Wang, L., Sun, J., Tong, X. (2008). Efficient estimation for the proportional hazards model with bivariate current status data. *Lifetime Data Analysis* 14:134–153.