

Statistica Sinica Preprint No: SS-2021-0411

Title	Generalized Odds Rate Frailty Models for Current Status Data with Informative Censoring
Manuscript ID	SS-2021-0411
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0411
Complete List of Authors	Yang Xu, Shishun Zhao, Tao Hu and Jianguo Sun
Corresponding Authors	Tao Hu
E-mails	hutao@cnu.edu.cn
Notice: Accepted version subject to English editing.	

Generalized Odds Rate Frailty Models for Current Status Data with Informative Censoring

Yang Xu^a, Shishun Zhao^a, Tao Hu^{b*} and Jianguo Sun^c

*a. Center for Applied Statistical Research and College of Mathematics,
Jilin University, Changchun, China*

b. School of Mathematical Sciences, Capital Normal University, Beijing, China

c. Department of Statistics, University of Missouri, Columbia, MO, USA

Abstract: Current status data occur in many areas and their analysis has recently attracted a great deal of attention. In this paper, we consider regression analysis of current status data in the presence of informative censoring, for which most of the existing methods either apply only to limited situations or are computationally unstable. Corresponding to this, we propose a new sieve maximum likelihood estimation procedure under the class of semiparametric, generalized odds rate frailty models, and in the method, the latent variable is employed to describe the informative censoring or relationship between the failure time of interest and the censoring time. For the determination of the proposed estimators, a novel EM algorithm is developed, and the asymptotic consistency and normality of the proposed estimators are established. An extensive simulation study is conducted

*Corresponding author. E-mail: hutaomath@foxmail.com

and indicates that the proposed method performs well in practical situations. In addition, the method is applied to a set of real data arising from a tumorigenicity experiment.

Key words and phrases: EM algorithm, Generalized odds rate frailty models, Informative censoring, Sieve approach.

1. Introduction

Current status data often occur in many areas, including clinical studies, epidemiology researches and sociological surveys. By current status data, we usually mean that in a failure time study, each study subject is observed only once and thus the failure time of interest is known only to be either smaller or larger than the observation time instead of being observed exactly. In other words, each observation is either left- or right-censored (Sun (2006); Jewell and Emerson (2013)). In this paper, we will discuss regression analysis of such data in the presence of dependent or informative censoring, meaning that the failure time of interest and the observation time may be correlated.

One area that often produces current status data with informative censoring is tumorigenicity experiments concerning the time to tumor onset. For the situation, it is usually the case that only the death or sacrifice time

of an animal is observed and one only knows the presence or absence of a tumor at the observation time. Thus only current status data are available on the tumor onset time. It is well-known that if the tumor is between lethal and non-lethal as often the case, the tumor onset time and the death time are correlated. A specific example of this on lung tumor will be discussed below in Section 6. A great deal of literature has been developed specifically for the analysis of tumorigenicity experiments, but most of the existing methods are parametric approaches (Dinse and Lagakos (1983); Lagakos and Louis (1988)).

The analysis of current status data has recently attracted a great deal of attention and this is especially the case for regression analysis of current status data with independent or non-informative censoring. For example, Huang (1996), Rossini and Tsiatis (1996) and Lin et al. (1998) discussed the problem under the proportional hazards (PH) model, the proportional odds (PO) model and the additive hazards model, respectively. Chen et al. (2009) also considered the problem but for multivariate situations. Note that as pointed out by Shiboski (1998) and others, in the presence of informative censoring, the analysis that ignores it could yield biased or misleading results or conclusions.

Some semiparametric methods have been proposed for regression analy-

sis of current status data with informative censoring and the existing methods can be generally classified into two types. One is the copula model-based methods, which use a copula function to describe the relationship between the failure time of interest and the observation time, and the other is the frailty-based approaches. Examples of the former include the methods given in Ma et al. (2015) and Zhao et al. (2015), who discussed the situations where the failure time of interest marginally follows the PH model or the additive hazards model, respectively. Du et al. (2019), Xu et al. (2019) and Xu et al. (2020) investigated and developed similar estimation procedures for the cases when the failure time of interest marginally follows a class of generalized probit models, a class of linear transformation models or the accelerated failure time model, respectively. Note that as pointed out by Ma et al. (2015) and others, a drawback of the copula model-based approach is that it is usually difficult or impossible to estimate the copula function and association parameter without imposing some strong assumptions.

The frailty-based methods employ latent variables to describe the association between the failure time of interest and the observation time. Among others, Zhang et al. (2005) and Li et al. (2017) proposed such methods under the assumption that the failure time of interest follows the additive hazards frailty model or the PH frailty model, respectively. It is well known

that either of the two models may not provide the appropriate fitting to the data since they are individual models. Corresponding to this, Chen et al. (2012) investigated the use of the transformation frailty model with a piecewise constant baseline hazard function and developed an EM algorithm. However, their EM algorithm can be less efficient and needs large computational effort.

The class of generalized odds rate (GOR) models has recently attracted a great deal of attention and includes both the PH and PO models as special cases. For example, Scharfstein et al. (1998) and Banerjee et al. (2007) discussed the fitting of the GOR model to right-censored failure time data. Also Zhou et al. (2017) proposed an EM algorithm for the maximum likelihood estimation of the GOR model based on interval-censored failure time data, meaning that the failure time of interest is observed only to belong to some intervals (Sun (2006)). However, all of the existing methods only apply to the situation of independent censoring. Corresponding to this, in the following, we will consider a class of GOR frailty models for regression analysis of current status data with informative censoring and propose a sieve maximum likelihood estimation procedure (Shen and Wong (1994); Shen (1997)). In particular, we will employ Gamma-Poisson latent variables to develop a novel and efficient EM algorithm.

The remainder of the paper is organized as follows. We will first introduce some notation and the GOR frailty model and then describe the resulting likelihood function in Section 2. The proposed estimation approach will be presented in Section 3 along with the development of the novel EM algorithm for the implementation of the proposed estimators. In the method, I -spline functions will be used to approximate the unknown functions involved, and Section 4 will establish the asymptotic properties of the proposed estimators. In Section 5, we will present some results from an extensive simulation study conducted to assess the finite sample performance of the proposed approach, and they indicate that it works well in practice. Finally we apply the proposed method to a tumorigenicity experiment in Section 6 and give some discussion and concluding remarks in Section 7.

2. Assumptions, Models and the Likelihood Function

Consider a failure time study consisting of n independent subjects. For subject i , let T_i denote the failure time of interest and X_i be the corresponding p -dimensional vector of covariates. Also for subject i , let C_i and C_i^c denote the two times related to the observation on T_i , where C_i may be related to T_i and C_i^c is independent of T_i . Suppose that each subject is observed

only at time $\tilde{C}_i = \min(C_i, C_i^c)$. That is, T_i is either left- or right-censored and we have only current status data on the T_i 's. In the tumorigenicity experiment example above, C_i represents the death time and C_i^c the sacrifice time. Define $\Delta_i = I(\tilde{C}_i = C_i)$, the administrative censoring indicator, and $\delta_i = I(T_i \leq \tilde{C}_i)$, the observed censoring indicator. Then the observed data have the form $\mathbf{O} = \{\mathbf{O}_i = (\tilde{C}_i, \Delta_i, \delta_i, X_i), i = 1, \dots, n\}$.

To describe the covariate effects and the relationship between T_i and C_i , suppose that there exists a latent variable b_i with mean one and variance η . Also suppose that given X_i and the random effect b_i , T_i follows the GOR frailty model with the cumulative hazard function given by

$$\Lambda_T(t|X_i, b_i) = G_r(\Lambda_1(t) \exp(X_i' \beta_1) b_i). \quad (2.1)$$

In the above, $\Lambda_1(t)$ denotes an unknown baseline cumulative hazard function, β_1 is a p -dimensional vector of regression parameters, and $G_r(\cdot)$ is a prespecified increasing transformation function indexed by a nonnegative argument r . It is easy to see that the model above includes many commonly used models as special cases. For example, by setting $G_0(x) = x$ and $G_r(x) = \log(1 + rx)/r$ with $r = 1$, model (2.1) gives the PH frailty model and the PO frailty model, respectively. Under model (2.1) with $G_r(x) = \log(1 + rx)/r$, the survival function of T_i given X_i and b_i can be

written as

$$S_T(t) = \begin{cases} \exp\{-\Lambda_1(t)e^{X_i'\beta_1}b_i\} & r = 0 \\ [1 + r\Lambda_1(t)e^{X_i'\beta_1}b_i]^{-\frac{1}{r}} & r > 0 \end{cases}.$$

In practice, covariates may have effects on C_i too. To describe this, assume that given X_i and b_i , the cumulative hazard function of C_i has the form

$$\Lambda_C(t|X_i, b_i) = \Lambda_2(t) \exp(X_i'\beta_2) b_i, \quad (2.2)$$

where $\Lambda_2(t)$ and β_2 are defined as $\Lambda_1(t)$ and β_1 , respectively. That is, C_i follows the PH frailty model. Let $S_C(t) = \exp\{-\Lambda_2(t) \exp(X_i'\beta_2) b_i\}$ and $f_C(t) = d\Lambda_2(t) \exp(X_i'\beta_2) b_i S_C(t)$, the survival and density functions of C_i given X_i and b_i , respectively, and assume that T_i and C_i are independent given b_i . Then the likelihood function of $\theta = (\beta_1', \beta_2', \eta, \Lambda_1, \Lambda_2)'$ can be written as

$$\begin{aligned} L_n(\theta|\mathbf{O}) &= \prod_{i=1}^n \left\{ \int_0^\infty [1 - S_T(\tilde{C}_i)] f_C(\tilde{C}_i) f(b_i; \eta) db_i \right\}^{\delta_i \Delta_i} \\ &\times \left\{ \int_0^\infty S_T(\tilde{C}_i) f_C(\tilde{C}_i) f(b_i; \eta) db_i \right\}^{(1-\delta_i)\Delta_i} \left\{ \int_0^\infty [1 - S_T(\tilde{C}_i)] S_C(\tilde{C}_i) f(b_i; \eta) db_i \right\}^{\delta_i(1-\Delta_i)} \\ &\times \left\{ \int_0^\infty S_T(\tilde{C}_i) S_C(\tilde{C}_i) f(b_i; \eta) db_i \right\}^{(1-\delta_i)(1-\Delta_i)}, \quad (2.3) \end{aligned}$$

where $f(\cdot; \eta)$ denotes the density function of the b_i 's, which is assumed to be known up to η .

To estimate $\boldsymbol{\theta}$, it is apparent that one natural method is to directly maximize the likelihood function given above. However, it is easy to see that this may be difficult and not reliable due to the complex structure of the likelihood function. To address this, we will first introduce a sieve approximation and then develop an efficient EM algorithm by incorporating Gamma-Poisson latent variables in the maximization process to solve the problems faced in the previous approaches.

3. Sieve Maximum Likelihood Estimation

Now we consider the estimation of the regression parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as well as others. As discussed above, it may not be easy to deal with unknown functions $\Lambda_1(t)$ and $\Lambda_2(t)$ based on the likelihood function $L_n(\boldsymbol{\theta}|\mathbf{O})$. For this, by following Ramsay (1988) and others, we employ the sieve approach to approximate $\Lambda_1(t)$ by using I -splines. For the function $\Lambda_2(t)$, it is apparent that one could use the same method as above but on the other hand, it will be seen below that it will be much easy to estimate it directly. Let Θ denote the parameter space of $\boldsymbol{\theta}$ and define the sieve space

$$\Theta_n = \{\boldsymbol{\theta}_n = ((\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \eta)', \Lambda_{1n}, \Lambda_2)\} = \mathcal{B} \otimes \mathcal{M}_n^1 \otimes \mathcal{M}^2,$$

where \mathcal{B} is compact subset of R^{2p+1} ,

$$\mathcal{M}_n^1 = \left\{ \Lambda_{1n}(t) = \sum_{l=1}^{K_n} \gamma_l I_l(t), M_1 \geq \gamma_l \geq 0, l = 1, \dots, K_n \right\},$$

$$\mathcal{M}^2 = \{ \Lambda_2(t) : 1/M_2 \leq \Lambda_2(t) \leq M_2 \},$$

with M_1 and M_2 being some positive constants. In the above, the I_l 's are nondecreasing integrated spline basis functions with each ranging from 0 to 1, and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{K_n})$ are nonnegative coefficients to ensure the monotonicity of $\Lambda_{1n}(t)$. The number of the spline basis functions, K_n , is equal to the number of interior knots plus the degree.

As mentioned above, even with the approximation, the maximization of the likelihood function is still not easy due to its complex form and the involvement of the latent variables b_i 's. To deal with this, a typical approach is to develop a standard EM algorithm as Chen et al. (2012) but it may not be stable and efficient. To overcome these issues, by following McMahan et al. (2013) and Li et al. (2017), we will employ the Gamma-Poisson data augmentation, which can greatly simplify the computing burden and yield stable and efficient estimators. To present the developed EM algorithm, first note that if the b_i 's were known, one could have the pseudo-complete data likelihood function as

$$L(\boldsymbol{\theta} | \mathbf{O}, \mathbf{b}) = \prod_{i=1}^n [1 - S_T(\tilde{C}_i)]^{\delta_i} S_T(\tilde{C}_i)^{1 - \delta_i} [d\Lambda_2(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_2) b_i]^{\Delta_i} S_C(\tilde{C}_i) f(b_i; \boldsymbol{\eta})$$

conditional on $\mathbf{b} = (b_1, \dots, b_n)$.

Let ϕ be a random variable following the Gamma distribution $\Gamma(1/r, r)$ with $r > 0$. Then the survival function of T_i can be rewritten as

$$S_T(t) = \left[1 + r\Lambda_1(t)e^{X_i'\beta_1 b_i}\right]^{-\frac{1}{r}} = \int_0^\infty \exp[-\phi\Lambda_1(t)e^{X_i'\beta_1 b_i}]f(\phi; r)d\phi.$$

Note that $\lim_{r \rightarrow 0} [1 + r\Lambda_1(t)e^{X_i'\beta_1 b_i}]^{-\frac{1}{r}} = \exp\{-\Lambda_1(t)e^{X_i'\beta_1 b_i}\}$, meaning that one can think that ϕ degenerates into a constant 1 for $r = 0$. Thus one could consider the new pseudo-complete data likelihood function

$$\begin{aligned} L_1(\boldsymbol{\theta}|\mathbf{O}, \mathbf{b}, \boldsymbol{\phi}, r) &= \prod_{i=1}^n \left\{1 - \exp\left[-\phi_i\Lambda_1(\tilde{C}_i) \exp(X_i'\beta_1) b_i\right]\right\}^{\delta_i} \\ &\times \left\{\exp\left[-\phi_i\Lambda_1(\tilde{C}_i) \exp(X_i'\beta_1) b_i\right]\right\}^{1-\delta_i} \left[d\Lambda_2(\tilde{C}_i) \exp(X_i'\beta_2) b_i\right]^{\Delta_i} \\ &\times \exp\left[-\Lambda_2(\tilde{C}_i) \exp(X_i'\beta_2) b_i\right] f(b_i; \eta) f(\phi_i; r) \end{aligned}$$

based on \mathbf{b} and $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)$, where $\phi_i \stackrel{\text{iid}}{\sim} \Gamma(1/r, r)$ for $i = 1, 2, \dots, n$.

Let $\{Z_i, i = 1, \dots, n\}$ be a set of independent Poisson random variable with means $\phi_i\Lambda_1(\tilde{C}_i) \exp(X_i'\beta_1) b_i$ given ϕ_i and b_i . It is easy to show that the likelihood function $L_1(\boldsymbol{\theta}|\mathbf{O}, \mathbf{b}, \boldsymbol{\phi}, r)$ can be equivalently written as

$$\begin{aligned} L_1(\boldsymbol{\theta}|\mathbf{O}, \mathbf{b}, \boldsymbol{\phi}, r) &= \prod_{i=1}^n P(Z_i > 0)^{\delta_i} P(Z_i = 0)^{1-\delta_i} \left[d\Lambda_2(\tilde{C}_i) \exp(X_i'\beta_2) b_i\right]^{\Delta_i} \\ &\times \exp\left[-\Lambda_2(\tilde{C}_i) \exp(X_i'\beta_2) b_i\right]. \end{aligned}$$

This suggests that one should consider the pseudo complete data likelihood function

$$L_2(\boldsymbol{\theta}|\mathbf{O}, \mathbf{b}, \boldsymbol{\phi}, \mathbf{Z}, r) = \prod_{i=1}^n \left[\frac{1}{Z_i!} \left(\phi_i \Lambda_1(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) b_i \right)^{Z_i} e^{-\phi_i \Lambda_1(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) b_i} \right] \\ \times \left[d\Lambda_2(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_2) b_i \right]^{\Delta_i} \exp \left[-\Lambda_2(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_2) b_i \right] f(b_i; \eta) f(\phi_i; r)$$

based on $\mathbf{b}, \boldsymbol{\phi}$ and $\mathbf{Z} = (Z_1, \dots, Z_n)$ and with the constraint $\delta_i I(Z_i > 0) + (1 - \delta_i) I(Z_i = 0) = 1$ for $i = 1, \dots, n$.

Finally note that one can decompose Z_i as the sum of K_n independent latent variables Z_{il} s with Z_{il} following the Poisson distribution with mean $\phi_i \gamma_l I_l(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) b_i$ conditional on ϕ_i and b_i , $l = 1, \dots, K_n$. Thus we propose to base the developed EM algorithm on the pseudo complete data likelihood function

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^n \left\{ \prod_{l=1}^{K_n} \frac{1}{Z_{il}!} \left(\phi_i \gamma_l I_l(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) b_i \right)^{Z_{il}} e^{-\phi_i \gamma_l I_l(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) b_i} \right\} \\ \times \{ d\Lambda_2(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_2) b_i \}^{\Delta_i} \exp \{ -\Lambda_2(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_2) b_i \} f(b_i; \eta) f(\phi_i; r)$$

based on $\{\mathbf{b}, \boldsymbol{\phi}, \mathbf{Z}_l = (Z_{1l}, \dots, Z_{nl}), l = 1, \dots, K_n\}$ and with the constraint $\delta_i I(\sum_{l=1}^{K_n} Z_{il} > 0) + (1 - \delta_i) I(\sum_{l=1}^{K_n} Z_{il} = 0) = 1$ for $i = 1, \dots, n$. Note that by integrating over $\mathbf{b}, \boldsymbol{\phi}, \mathbf{Z}_l$ s, $L_c(\boldsymbol{\theta})$ will reduce to the observed data likelihood function given in (2.3).

Let $l_c(\boldsymbol{\theta}) = \log L_c(\boldsymbol{\theta})$ and define $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, \eta)'$ with $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}'_1, \Lambda_{1n})'$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}'_2, \Lambda_2)'$. In the E-step of the EM algorithm, we need to determine

the conditional expectation of $l_c(\boldsymbol{\theta})$ with respect to all latent variables. Specifically, at the m th iteration, the conditional expectation of $l_c(\boldsymbol{\theta})$ can be expressed as the summation of the four parts

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}) = Q_1(\boldsymbol{\theta}_1; \boldsymbol{\theta}^{(m)}) + Q_2(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(m)}) + Q_3(\eta; \boldsymbol{\theta}^{(m)}) + Q_4(\boldsymbol{\theta}^{(m)}),$$

where

$$Q_1(\boldsymbol{\theta}_1; \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^n \sum_{l=1}^{K_n} \{\log(\gamma_l) + X_i' \boldsymbol{\beta}_1\} \hat{E}(Z_{il}) - \gamma_l I_l(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) \hat{E}(\phi_i b_i),$$

$$Q_2(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^n \Delta_i (\log d\Lambda_2(\tilde{C}_i) + X_i' \boldsymbol{\beta}_2) - \Lambda_2(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_2) \hat{E}(b_i),$$

$$Q_3(\eta; \boldsymbol{\theta}^{(m)}) = n\eta^{-1} \log \eta^{-1} - n \log \Gamma(\eta^{-1}) + \eta^{-1} \sum_{i=1}^n \{\hat{E}[\log(b_i)] - \hat{E}(b_i)\},$$

and $Q_4(\boldsymbol{\theta}^{(m)})$ denotes a function of $\boldsymbol{\theta}^{(m)}$ free of $\boldsymbol{\theta}$. In the above, $\hat{E}(Z_{il})$, $\hat{E}(\phi_i b_i)$, $\hat{E}(b_i)$ and $\hat{E}(\log b_i)$ denote the conditional expectations with respect to all latent variables.

For the calculation of the conditional expectations above, note that

$$\hat{E}(Z_i) = \Delta_i \Lambda_{1n}(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) E_{b_i} \left\{ \frac{b_i}{1 - S_T(\tilde{C}_i)} \right\}.$$

Using the fact $Z_{il} | Z_i \sim \text{Binomial} \left(Z_i, \frac{\phi_i \gamma_l I_l(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) b_i}{\phi_i \Lambda_{1n}(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) b_i} \right)$ and applying the iterated rule of expectations, we have that

$$\hat{E}(Z_{il}) = \frac{\gamma_l I_l(\tilde{C}_i)}{\Lambda_{1n}(\tilde{C}_i)} \hat{E}(Z_i) = \Delta_i \gamma_l I_l(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) E_{b_i} \left\{ \frac{b_i}{1 - S_T(\tilde{C}_i)} \right\},$$

and

$$\hat{E}(\phi_i b_i) = E_{b_i} \left\{ b_i S_T(\tilde{C}_i)^{r(1-\delta_i)} \left[\frac{1 - S_T(\tilde{C}_i)^{r+1}}{1 - S_T(\tilde{C}_i)} \right]^{\delta_i} \right\}.$$

It is clear that all conditional expectations above do not have closed forms and for this, by following others, we can approximate them by the Monte Carlo method. Specifically, the conditional expectation of any arbitrary function $h(b_i)$ and sufficiently large L can be calculated through

$$\begin{aligned} E_{b_i} \{h(b_i) \mid O_i, \hat{\boldsymbol{\theta}}^{(m)}\} &= \frac{\int_{-\infty}^{\infty} h(b_i) \Psi_i(b_i; \hat{\boldsymbol{\theta}}^{(m)}) f(b_i; \hat{\eta}^{(m)}) db_i}{\int_{-\infty}^{\infty} \Psi_i(b_i; \hat{\boldsymbol{\theta}}^{(m)}) f(b_i; \hat{\eta}^{(m)}) db_i} \\ &\approx \frac{L^{-1} \sum_{i=1}^L h(b_{il}) \Psi_i(b_{il}; \hat{\boldsymbol{\theta}}^{(m)})}{L^{-1} \sum_{i=1}^L \Psi_i(b_{il}; \hat{\boldsymbol{\theta}}^{(m)})}. \end{aligned}$$

In the above,

$$\Psi_i(b_i; \boldsymbol{\theta}) = [1 - S_T(\tilde{C}_i)]^{\delta_i} S_T(\tilde{C}_i)^{1-\delta_i} [d\Lambda_2(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_2) b_i]^{\Delta_i} S_C(\tilde{C}_i),$$

and $\{b_{il}, l = 1, \dots, L\}$ are generated from the density function $f(b_i; \eta^{(m)})$.

In the M-step of the EM algorithm, by setting $\partial Q_1(\boldsymbol{\theta}_1; \boldsymbol{\theta}^{(m)}) / \partial \gamma_l = 0$, we can calculate γ_l with the following closed-form expression

$$\gamma_l = \frac{\sum_{i=1}^n \hat{E}(Z_{il})}{\sum_{i=1}^n I_l(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) \hat{E}(\phi_i b_i)}, \quad l = 1, \dots, K_n. \quad (3.4)$$

By plugging the estimators above into $Q_1(\boldsymbol{\theta}_1; \boldsymbol{\theta}^{(m)})$, we can obtain the score functions for $\boldsymbol{\beta}_1$ as

$$\sum_{i=1}^n \sum_{l=1}^{K_n} \hat{E}(Z_{il}) \left\{ X_i - \frac{\sum_{i=1}^n I_l(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) \hat{E}(\phi_i b_i) X_i}{\sum_{i=1}^n I_l(\tilde{C}_i) \exp(X_i' \boldsymbol{\beta}_1) \hat{E}(\phi_i b_i)} \right\} = 0. \quad (3.5)$$

For the determination of the updated estimators of β_2 and $\Lambda_2(t)$, by treating $\Lambda_2(t)$ as a piecewise constant function between the uncensored observation times, one can obtain the following score function

$$\frac{\partial Q_2(\theta_2, \theta^{(m)})}{\partial \beta_2} = \sum_{i=1}^n \int_0^{\infty} \{X_i - \bar{X}(t, \beta_2)\} dN_i(t) = 0, \quad (3.6)$$

where $N_i(t) = I(\tilde{C}_i \leq t, \Delta_i = 1)$ and

$$\bar{X}(t, \beta_2) = \frac{\sum_{i=1}^n Y_i(t) \exp(X_i' \beta_2) \hat{E}(b_i) X_i}{\sum_{i=1}^n Y_i(t) \exp(X_i' \beta_2) \hat{E}(b_i)}$$

with $Y_i(t) = I(\tilde{C}_i \geq t)$. Thus, the estimator $\hat{\Lambda}_2^{(m+1)}$ can be updated by the following Breslow-type estimator

$$\hat{\Lambda}_2^{(m+1)}(t) = \sum_{i=1}^n \int_0^t \frac{dN_i(s)}{\sum_{i=1}^n Y_i(s) \exp(X_i' \beta_2^{(m+1)}) \hat{E}(b_i)}. \quad (3.7)$$

Finally, the updated estimator of η can be obtained by solving the score equation

$$\partial Q_3(\eta; \theta^{(m)}) / \partial \eta = 0.$$

By combining all steps above, the proposed EM algorithm can be summarized as follows.

Step 0. Select the initial estimate of $\theta^{(0)}$ such as $\beta_1^{(0)} = \beta_2^{(0)} = \mathbf{0}, \gamma^{(0)} = \mathbf{1}, \eta = 1$ and a pre-specified value of r ;

Step 1. At the $(m + 1)$ th iteration, generate the random sample $\{b_{il}, l = 1, \dots, L, i = 1, \dots, n\}$ from the density function $f(b_i; \eta^{(m)})$;

Step 2. Calculate the conditional expectations $\hat{E}(Z_{il})$, $\hat{E}(\phi_i b_i)$, $\hat{E}(b_i)$ and

$\hat{E}\{\log(f(b_i; \eta))\}$ based on $\boldsymbol{\theta}^{(m)}$;

Step 3. Update $\hat{\boldsymbol{\beta}}_1^{(m+1)}$ based on (3.5) by using the Newton-Raphson method

and then determine $\gamma_l^{(m+1)}$ by (3.4) for $l = 1, \dots, K_n$;

Step 4. Update $\hat{\boldsymbol{\beta}}_2^{(m+1)}$ based on (3.6) by using the Newton-Raphson method

and then determine $\hat{\Lambda}_2^{(m+1)}$ by (3.7);

Step 5. Calculate $\eta^{(m+1)}$ by solving $\partial Q_3(\eta; \boldsymbol{\theta}^{(m)})/\partial \eta = 0$;

Step 6. Repeat Steps 1 - 5 until the estimates between two consecutive

iterations is less than a pre-specified constant such as 0.0001.

For the implementation of the EM algorithm above, it is apparent that one needs to choose the degree of the splines and the number of interior knots. For this, one commonly used approach is to arrange the knots equally spaced or based on the quartiles and set the degree of the splines to be 2 or 3. An alternative is to try different values and apply the Akaike's information criterion (AIC) criterion defined as

$$AIC = -2l_n(\hat{\boldsymbol{\theta}}_n) + 2df_k$$

to choose the optimal model. In the above, $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2', \hat{\eta}, \hat{\Lambda}_{1n}, \hat{\Lambda}_2)'$ denotes the estimator defined above, $l_n(\hat{\boldsymbol{\theta}}_n) = \log L_n(\hat{\boldsymbol{\theta}}_n)$, and df_k denotes the

number of the parameters to be estimated. The same AIC criterion can also be used to choose the nonnegative argument r , which has been assumed to be known but may not be in practice. More specifically, one could perform the grid search by considering a sequence of values for r and choose the value that gives the smallest AIC value.

4. Asymptotic Properties

In this section, we discuss the asymptotic properties of the estimator proposed above and for this, let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \eta_0, \Lambda_{10}, \Lambda_{20})$ denote the true value of $\boldsymbol{\theta}$ and for two different $\check{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n \in \Theta_n$, define the norm

$$d(\check{\boldsymbol{\theta}}_n, \tilde{\boldsymbol{\theta}}_n) = \left(\|\check{\Lambda}_{1n} - \tilde{\Lambda}_{1n}\|_2^2 + \|\check{\Lambda}_2 - \tilde{\Lambda}_2\|_2^2 + \|\check{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1\|_E^2 + \|\check{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_2\|_E^2 + \|\check{\eta} - \tilde{\eta}\|_E^2 \right)^{1/2},$$

where $\|\cdot\|_2$ and $\|\cdot\|_E$ denote the L^2 and Euclidean norms, respectively. For the asymptotic properties, we need the following regularity conditions.

(C1) The maximum spacing of the knots satisfies $\tilde{\Delta} = \max_{2 \leq j \leq k_n+2} |t_j - t_{j-1}| = O(n^{-\nu})$ with $\nu \in (0, 0.5)$. Moreover, there exists a constant $M > 0$ such that $\tilde{\Delta}/\tilde{\delta} \leq M$ uniformly in n , where $\tilde{\delta} = \min_{2 \leq j \leq k_n+2} |t_j - t_{j-1}|$ and k_n is the number of interior knots.

(C2) The covariates X_i 's have a bounded support in R^p .

(C3) The true cumulative hazard functions, $\Lambda_{10}(t)$ and $\Lambda_{20}(t)$, are increasing functions. Also the κ th derivative of $\Lambda_{10}(t)$ is bounded and continuous, $\Lambda_{20}(t)$ satisfies $1/M_2 \leq \Lambda_{20}(t) \leq M_2$.

(C4) Define $\boldsymbol{\vartheta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \eta)'$ and

$$L_{(\boldsymbol{\vartheta}, y_1, y_2)}(\delta, \Delta, X) = \int_0^\infty [1 - S_T(\tilde{c}; \boldsymbol{\vartheta}, y_1)]^\delta S_T(\tilde{c}; \boldsymbol{\vartheta}, y_1)^{1-\delta} f_C(\tilde{c}; \boldsymbol{\vartheta}, y_2)^\Delta \\ \times S_C(\tilde{c}; \boldsymbol{\vartheta}, y_2)^{1-\Delta} f(b; \eta) db.$$

Suppose that there exists t^* in the support of the conditional distribution of \tilde{C} given $X = x$ for which there are $2p + 3$ different values (δ, Δ, X) such that if

$$\left(a'_1 \frac{\partial}{\partial \boldsymbol{\vartheta}} + a_2 \frac{\partial}{\partial y_1} + a_3 \frac{\partial}{\partial y_2} \right) \log L_{(\boldsymbol{\vartheta}, y_1, y_2)}(\delta, \Delta, X) \Big|_{(\boldsymbol{\vartheta}, y_1, y_2) = (\boldsymbol{\vartheta}_0, \Lambda_1(t^*), \Lambda_2(t^*))} = 0,$$

then $a_1 = \mathbf{0}, a_2 = a_3 = 0$, where $(\delta, \Delta) = (0, 0), (1, 0), (0, 1)$ or $(1, 1)$, $S_C(t; \boldsymbol{\vartheta}, y_2(t)) = \exp\{-y_2(t) \exp(X' \boldsymbol{\beta}_2) b\}$, $f_C(t; \boldsymbol{\vartheta}, y_2(t)) = dy_2(t) \exp(X' \boldsymbol{\beta}_2) b S_C(t; \boldsymbol{\vartheta}, y_2(t))$ and

$$S_T(t; \boldsymbol{\vartheta}, y_1(t)) = \begin{cases} \exp\{-y_1(t) e^{X' \boldsymbol{\beta}_1} b\} & r = 0 \\ [1 + r y_1(t) e^{X' \boldsymbol{\beta}_1} b]^{-\frac{1}{r}} & r > 0 \end{cases}.$$

(C5) The matrix Σ defined in the supplementary material is finite and positive definite.

Note that the conditions above are mild and usually satisfied in practical situations. In particular, Condition **(C1)** is the same as Condition 1 of Lu et al. (2007), and it is necessary for the construction of spline sieve space and the consistency of the spline likelihood-based estimators. The bounded support assumption stated in Condition **(C2)** for covariates is commonly used in the current status data literature and needed for the uniform convergence. See more detailed discussions on this in Van der Vaart and Wellner (1996). Condition **(C3)**, the smoothness assumption of the cumulative hazards functions, is standard in the nonparametric smoothing literature. It is easy to see that Condition **(C4)** is a sufficient condition for the identifiability of the parameters, which is common in existing literature (Li et al. (2001); Chang et al. (2007)). The following theorems give the asymptotic properties.

Theorem 1. Assume that Conditions **(C1)**-**(C4)** given above hold. Then as $n \rightarrow \infty$, we have that

$$d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(n^{-(1-v)/2} + n^{-\kappa v}).$$

Theorem 2. Assume that Conditions **(C1)**-**(C5)** given above hold. Then as $n \rightarrow \infty$, we have that

$$n^{1/2} \boldsymbol{\alpha}' \left((\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})', (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})', (\hat{\eta} - \eta_0) \right)' + \int_0^{\tau_c} g(t) d \left(\hat{\Lambda}_2(t) - \Lambda_{20}(t) \right) \xrightarrow{D} N(0, \Sigma),$$

where $\boldsymbol{\alpha}$ is any $(2p+1)$ -dimensional vector with $\|\boldsymbol{\alpha}\|_E \leq 1$, τ_c denotes the longest follow-up time, g is a function with bounded variation on $[0, \tau_c]$, and Σ is the semiparametric efficiency bound defined in the supplementary material.

The proof of the theorems above is sketched in the supplementary material. Note that Theorem 1 states that the proposed estimator is consistent and can achieve the optimal convergence rate. Theorem 2 says that the proposed estimators $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2$ and $\hat{\eta}$ are asymptotically efficient. In particular, by setting $g = 0$ in Theorem 2, we have that

$$n^{1/2} \boldsymbol{\alpha}'((\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})', (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})', (\hat{\eta} - \eta_0))' \xrightarrow{D} N(0, \Sigma_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \eta}).$$

It is apparent that for inference about $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and η , we need to estimate the asymptotic covariance matrix of the corresponding estimators. Since it would be quite difficult to derive the consistent estimator of $\Sigma_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \eta}$, we suggest to employ the nonparametric bootstrap method (Efron (1981)) as follows. We first repeatedly draw new datasets of sample size n with replacement from the original observed data \boldsymbol{O} for Q times, where Q is a prespecified positive integer. The newly resampled datasets are denoted by $\boldsymbol{O}^{(q)}$ for $q = 1, \dots, Q$. Let $\hat{\boldsymbol{\beta}}_1^{(q)}, \hat{\boldsymbol{\beta}}_2^{(q)}$ and $\hat{\eta}^{(q)}$ denote the maximum likelihood estimators of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and η based on the bootstrap sample $\boldsymbol{O}^{(q)}$.

Then one can estimate the covariance matrix or variance of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\eta}$ by the empirical covariance matrix or variance of the $\hat{\beta}_1^{(q)}$'s, $\hat{\beta}_2^{(q)}$'s and $\hat{\eta}^{(q)}$'s, respectively. The simulation study below suggests that this method works well for practical situations.

5. A Simulation Study

An extensive simulation study was conducted to assess the finite sample performance of the estimation procedure proposed in the previous sections. In the study, we considered two covariate X_{i1} and X_{i2} with the first covariate from the Bernoulli distribution with the success probability of 0.5 and the second covariate from the uniform distribution over $(0, 1)$. The latent variables b_i 's were generated from the gamma distribution with mean one and variance $\eta = 0.4$. For the generation of the observed data, it was assumed that the T_i 's and C_i 's follow models (2.1) and (2.2), respectively, with $\Lambda_1(t) = \Lambda_2(t) = 0.05t^2$ or $\Lambda_1(t) = 0.5 \log(1 + t)$ and $\Lambda_2(t) = 0.05t^2$. The independent observation times C_i^c 's were set to be the constant τ_c , which was chosen to yield the desired censoring percentage. The results given below are based on 500 replications with $Q = 50$, $n = 200$ and $r = 0, 0.5, 1$ or 2.

Tables 1 and 2 present the results on the estimation of the parameters

given by the proposed method with the same baseline function situation and various true values. The results include the the estimated bias (Bias) given by the average of the estimates minus the true value, the sample standard error (SSE) of the estimates, the average of the estimated standard errors (SEE), and the 95% empirical coverage probability (CP). Here for the monotone splines approximation, following Wang et al. (2016), we used the 5 equally spaced knots in terms of percentiles and the order of 3 for monotone splines.

One can see from Tables 1 and 2 that the proposed maximum likelihood estimators seem to be unbiased and the bootstrap variance estimates are appropriate. In addition, the normal approximation to the distribution of the estimators appears to be reasonable because all empirical coverage probabilities are close the nominal value. The results given in Tables 3 and 4 were obtained with different baseline functions but the other set-ups being the same as in Tables 1 and 2 on the estimation of regression parameters. It is easy to see that they are similar to those presented in Tables 1 and 2 and again suggest that the proposed method performs well for the situations considered.

Note that the proposed estimation procedure assumes that the distribution of the latent variables b'_i s is known up to the parameter η and it

is clear that this may not be true in practice. Thus a question of interest is the robustness of the proposed estimation procedure to the distribution. To assess this, we repeated the study giving the results reported in Table 1 except that we generated b_i 's from the **log-normal distribution with mean one and variance $\eta = 0.4$** . The estimation was still carried out based on the gamma distribution and the obtained results are presented in Table 5. One can see that the proposed estimators seem to perform well, and the results suggest that the estimation procedure is robust to the misspecification of the distribution of the latent variable. **We also considered some other set-ups including different sample sizes and different numbers of knots and orders for monotone splines. As expected, the performance became better when the sample size increased, and it seems that the results are quite stable with respect to different numbers of knots and orders.**

6. An Application

In this section, we apply the methodology proposed in the previous sections to the tumorigenicity experiment described in Sun (2006) among others. The data consist of 144 RFM mice and only current status data are available on the time to tumor onset, the variable of interest, since the tumor status was examined only at death. Also the experiment involves two treatments,

germ-free environment (48 mice) and conventional environment (96 mice). Among the animals in the two groups, 35 and 27 mice were observed to have tumors at the time of death, respectively, and since the lung tumor is between lethal and non-lethal, we have dependent or informative censoring. One objective of the study is to compare the tumor growth rates between the two different environments or treatments.

To apply the proposed estimation procedure, let T_i denote the time to tumor onset and define $X_i = 1$ if the i th mice was in germ-free environment and 0 otherwise. As discussed above, for the analysis, one needs to select the degree of splines, the number of knots of splines and the nonnegative argument r , and for these, we used the 3-dimensional grid search based on the AIC values. More specifically, we considered the degree of the monotone splines being 2, 3 or 4 and the number of interior knots varying from 3 to 8 to provide sufficient model flexibility and less computational burden. For r , we considered the values ranging from 0 to 2 with increments of 0.1, and the optimal frailty model was given by 2 degrees of splines with 4 interior knots and $r = 0.4$.

Table 6 gives the obtained results on the estimation of covariates effects, including the estimated effects, the estimated standard errors and the p -values for testing no covariate effect, under the optimal model. For

comparison, we also include the results based on 3 and 4 degrees of splines with the selected r and the number of interior knots as well as the corresponding results obtained under the PH frailty model and the PO frailty model, respectively. One can see from Table 6 that all results seem to be consistent. They suggest that the animals in the germ-free environment had significantly longer survival times than those in the conventional environment but the animals in the two groups seem to have similar risks of developing tumors. In addition, the results indicate that the tumor onset time and the death time were significantly correlated.

Also for comparison, we reanalyzed the data by imposing the non-informative censoring or independence assumption between the tumor onset time and the death time through setting $b_i = 1$ in the models (2.1) and (2.2) and include the obtained results in Table 6. One can see that they suggest that the animals in the germ-free environment had significantly higher rate of tumor growth than those in the conventional environment. In other words, one would get different conclusions if missing or ignoring the informative censoring.

7. Discussion and Concluding Remarks

This paper proposed a sieve semiparametric maximum likelihood estimation procedure for regression analysis of current status data in the presence of informative censoring. In the method, the GOR frailty model was employed to describe covariate effects as well as the association between the failure time of interest and the observation time, and for the implementation of the methodology, a novel EM algorithm was developed. In particular, the observed data were augmented by introducing some latent variables and the proposed algorithm is computationally efficient. In addition, the proposed estimators were shown to be consistent and asymptotically normal, and the numerical results suggested that the presented estimation procedure seems to work well for practical situations.

As mentioned above, to address informative censoring, an alternative to the proposed frailty-based approach is the copula model-based approach. There exist a couple of differences between the two methods. One is that unlike the latter, the former gives unique estimators, while the latter can have an identifiability issue. In particular, the correlation parameters are only included in the joint distribution function for the latter, while the correlation parameters are included in both marginal and joint distribution functions for the latter.

In the method proposed above, I -spline functions have been employed to approximate the unknown functions involved and to reduce the estimation of an infinite-dimensional function to the estimation of finite parameters. As an alternative, one can use other smooth functions such as B -spline functions and the method can be developed similarly. Another assumption used above is that the observation time C follows the PH frailty model and one advantage of it is that it can reduce the computation burden. Instead one could use some other models such as the PO frailty model or the GOR frailty model to model covariate effects on C and the proposed approach can still be applied.

Note that although the proposed approach applies to a general class of GOR frailty models characterized by the nonnegative argument r , it has been assumed that r is known. It is apparent that it would be useful to develop some simultaneous estimation procedures for r . However, this is usually not possible without some extra assumptions or information. An alternative is to choose r based on some selection criterion to give an indication for a preferred model as illustrated in the real data analysis. Also note that GOR frailty models do not include the additive hazards frailty model as a special case. Furthermore, it remains challenging to develop statistical methods to assess the goodness-of-fit of each model and compare

the appropriateness of both models.

With respect to the distribution of the frailty, it is worth noting that in the numerical study, we have focused on the gamma distribution as it has a large left tail and thus is more appropriate for strong late dependence (Hougaard (1995)), which is usually the case for tumorigenicity experiments. On the other hand, the numerical results suggested that the proposed estimation procedure seems to be robust to the misspecification of the distribution of the b_i 's. Of course, one may need to be careful about this as only limited studies were performed.

More work remains to be done or there exist several directions for future research. One is that in the previous sections, we have focused on current status data and sometimes one may face informative interval-censored failure time data, which include current status data as a special case (Sun (2006)). Thus it would be useful to generalize the proposed method to the latter situation. Another situation that may occur in practice is that instead of the univariate informative current status data, one may encounter multivariate informative current status data and hence it would be useful to generalize the proposed method to allow multivariate failure times of interest. Model checking is apparently also important for the models assumed above and it does not seem to exist any literature discussing this.

Supplementary Materials

All technical proofs are given in the supplementary material.

Acknowledgements

The authors would like to thank the editor and two referees for their many comments and suggestions that greatly improved the paper. This research was funded by the Beijing Natural Science Foundation Z210003 and National Nature Science Foundation of China (Grant Nos: 12171328, 11971064). Shishun Zhao's study was supported by National Natural Science Foundation of China (NSFC) (12071176). **The R code for the proposed method can be downloaded at <https://github.com/xymath/EM.Iter>.**

References

- Banerjee, T., Chen, M. H., Dey, D. K. and Kim, S. (2007). Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Analysis* **13**, 241-260.
- Chang, I. S., Wen, C. C. and Wu, Y. J. (2007). A profile likelihood theory for the correlated gamma-frailty model with current status family data. *Statistica Sinica* **17**, 1023-1046.
- Chen, C. M., Lu, T. F. C., Chen, M. H. and Hsu, C. M. (2012). Semiparametric transformation models for current status data with informative censoring. *Biometrical Journal* **54**, 641-656.

- Chen, M. H., Tong, X. and Sun, J. (2009). A frailty model approach for regression analysis of multivariate current status data. *Statistics in Medicine* **28**, 3424-3436.
- Dinse, G. E. and Lagakos, S. W. (1983). Regression analysis of tumour prevalence data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **32**, 236-248.
- Du, M., Hu, T. and Sun, J. (2019). Semiparametric probit model for informative current status data. *Statistics in Medicine* **38**, 2219-2227.
- Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76**, 312-319.
- Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis* **1**, 255-273.
- Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics* **24**, 540-568.
- Jewell N. P. and Emerson R. (2013). Current status data: An illustration with data on avalanche victims. In *Handbook of Survival Analysis*. Boca Raton, Florida: Chapman & Hall/CRC, pp. 391C412.
- Lagakos, S. W., and Louis, T. A. (1988). Use of tumour lethality to interpret tumorigenicity experiments lacking cause-of-death data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **37**, 169-179.
- Li, C. S., Taylor, J. M. and Sy, J. P. (2001). Identifiability of cure models. *Statistics and Probability Letters* **54**, 389-395.

- Li, S., Hu, T., Wang, P. and Sun, J. (2017). Regression analysis of current status data in the presence of dependent censoring with applications to tumorigenicity experiments. *Computational Statistics and Data Analysis* **110**, 75-86.
- Li, S., Hu, T., Zhao, S. and Sun, J. (2020). Regression analysis of multivariate current status data with semiparametric transformation frailty models. *Statistica Sinica* **30**, 1117-1134.
- Lin, D., Oakes, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85**, 289-298.
- Liu, Y., Hu, T. and Sun, J. (2017). Regression analysis of current status data in the presence of a cured subgroup and dependent censoring. *Lifetime Data Analysis* **23**, 626-650.
- Lu, M., Zhang, Y. and Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* **84**, 705-718.
- Ma, L., Hu, T. and Sun, J. (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* **102**, 731-738.
- McMahan, C. S., Wang, L. and Tebbs, J. M. (2013). Regression analysis for current status data using the EM algorithm. *Statistics in Medicine* **32**, 4452-4466.
- Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**, 425-441.
- Rossini, A. J. and 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.

- Scharfstein, D. O., Tsiatis, A. A. and Gilbert, P. B. (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime Data Analysis* **4**, 355-391.
- Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics* **25**, 2555-2591.
- Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics* **22**, 580-615.
- Shiboski, S. C. (1998). Generalized additive models for current status data. *Lifetime Data Analysis* **4**, 29-50.
- Sun J. (2006). *The statistical analysis of interval-censored failure time data*. New York: Springer.
- Van der Vaart, A. W. and Wellner, J. (1996) *Weak convergence and empirical processes*. New York: Springer.
- Wang, L., McMahan, C. S., Hudgens, M. G. and Qureshi, Z. P. (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics* **72**, 222-231.
- Xu, D., Zhao, S., Hu, T., Yu, M. and Sun, J. (2019). Regression analysis of informative current status data with the semiparametric linear transformation model. *Journal of Applied Statistics* **46**, 187-202.
- Xu, D., Zhao, S. and Sun, J. (2020). Regression analysis of dependent current status data with the accelerated failure time model. *Communications in Statistics-Simulation and Compu-*

tation, 1-9.

Zhang, Z., Sun, J. and Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine* **24**, 1399-1407.

Zhao, S., Hu, T., Ma, L., Wang, P. and Sun, J. (2015). Regression analysis of informative current status data with the additive hazards model. *Lifetime Data Analysis* **21**, 241-258.

Zhou, J., Zhang, J. and Lu, W. (2017). An Expectation Maximization algorithm for fitting the generalized odds-rate model to interval censored data. *Statistics in Medicine* **36**, 1157-1171.

Table 1: Simulation results on regression parameters based on the simulated data with same baseline functions.

r	Par.	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
		$(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}) = (0, 0, 0, 0)$				$(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}) = (0.2, 0.2, 0.2, 0.2)$			
0	β_{11}	0.007	0.274	0.280	95.2	0.027	0.278	0.278	95.4
	β_{12}	0.029	0.481	0.497	94.8	0.033	0.496	0.480	94.8
	β_{21}	0.006	0.264	0.262	93.0	0.062	0.258	0.255	93.4
	β_{22}	0.017	0.504	0.506	94.4	0.032	0.477	0.470	95.4
	η	0.005	0.128	0.151	96.4	0.017	0.126	0.116	93.8
0.5	β_{11}	-0.009	0.309	0.320	95.6	0.014	0.329	0.318	93.4
	β_{12}	0.019	0.562	0.565	93.8	0.072	0.560	0.551	94.2
	β_{21}	-0.009	0.259	0.257	94.4	0.044	0.279	0.251	91.6
	β_{22}	-0.014	0.482	0.494	93.0	0.043	0.492	0.469	93.6
	η	-0.018	0.112	0.138	95.4	-0.006	0.122	0.113	92.8
1	β_{11}	0.004	0.347	0.353	94.8	0.013	0.360	0.360	93.4
	β_{12}	-0.005	0.624	0.630	95.8	0.038	0.607	0.620	93.4
	β_{21}	-0.004	0.253	0.249	93.6	0.045	0.258	0.248	93.4
	β_{22}	0.009	0.478	0.481	94.2	0.031	0.484	0.462	94.0
	η	-0.034	0.114	0.129	93.2	-0.016	0.115	0.111	92.6
2	β_{11}	-0.001	0.442	0.431	93.2	0.026	0.431	0.432	93.6
	β_{12}	-0.031	0.782	0.751	93.2	0.051	0.783	0.750	93.2
	β_{21}	-0.009	0.253	0.242	92.8	0.054	0.274	0.243	92.2
	β_{22}	0.015	0.483	0.457	92.0	0.035	0.486	0.452	92.6
	η	-0.060	0.110	0.121	91.2	-0.049	0.102	0.106	92.4

Table 2: Simulation results on regression parameters based on the simulated data with same baseline functions.

r	Par.	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP	
		$(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})=(0.2, 0.2, -0.2, -0.2)$				$(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})=(-0.2, -0.2, -0.2, -0.2)$				
0	β_{11}	0.005	0.279	0.268	94.4	-0.033	0.278	0.280	93.4	
	β_{12}	0.025	0.453	0.470	95.8	-0.005	0.472	0.497	95.6	
	β_{21}	-0.053	0.240	0.261	96.8	-0.063	0.250	0.267	96.4	
	β_{22}	-0.026	0.510	0.499	94.2	-0.024	0.485	0.498	95.4	
	η	-0.029	0.085	0.119	95.0	-0.032	0.137	0.165	93.8	
0.5	β_{11}	-0.009	0.309	0.320	95.6	-0.006	0.323	0.326	96.0	
	β_{12}	0.019	0.562	0.565	93.8	-0.009	0.556	0.570	95.6	
	β_{21}	-0.009	0.259	0.257	94.4	-0.057	0.258	0.263	95.0	
	β_{22}	-0.014	0.482	0.494	93.0	-0.042	0.480	0.495	95.2	
	η	-0.018	0.112	0.138	95.4	-0.032	0.125	0.159	92.6	
1	β_{11}	-0.010	0.357	0.350	93.6	-0.003	0.362	0.361	92.4	
	β_{12}	-0.029	0.605	0.612	95.0	-0.006	0.627	0.630	95.2	
	β_{21}	-0.037	0.260	0.254	94.2	-0.027	0.253	0.255	95.0	
	β_{22}	-0.003	0.497	0.482	92.8	-0.012	0.513	0.481	93.2	
	η	-0.053	0.091	0.119	93.2	-0.051	0.117	0.141	92.2	
2	β_{11}	0.022	0.442	0.424	94.0	-0.002	0.418	0.433	96.4	
	β_{12}	0.022	0.746	0.746	93.6	-0.011	0.742	0.758	94.2	
	β_{21}	-0.032	0.248	0.247	95.0	-0.019	0.252	0.245	94.8	
	β_{22}	-0.068	0.491	0.462	93.6	-0.047	0.474	0.460	92.2	
	η	-0.059	0.090	0.114	90.8	-0.069	0.102	0.130	90.0	

Table 3: Simulation results on regression parameters based on the simulated data with different baseline functions.

r	Par.	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
		$(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}) = (0, 0, 0, 0)$				$(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}) = (0.2, 0.2, 0.2, 0.2)$			
0	β_{11}	-0.007	0.244	0.259	95.4	0.042	0.245	0.255	94.1
	β_{12}	0.032	0.438	0.456	95.2	0.076	0.423	0.443	95.5
	β_{21}	0.006	0.254	0.259	95.0	0.062	0.262	0.255	92.0
	β_{22}	-0.010	0.453	0.486	94.8	0.042	0.501	0.460	91.4
	η	0.012	0.174	0.213	95.2	0.034	0.230	0.202	90.8
0.5	β_{11}	-0.011	0.293	0.301	94.2	0.046	0.300	0.303	95.3
	β_{12}	-0.018	0.509	0.527	94.4	0.033	0.514	0.525	93.6
	β_{21}	0.019	0.264	0.258	94.6	0.049	0.267	0.260	93.6
	β_{22}	-0.005	0.471	0.483	95.0	0.034	0.513	0.471	91.2
	η	-0.006	0.187	0.204	94.0	-0.049	0.214	0.195	91.6
1	β_{11}	-0.016	0.327	0.338	94.6	0.030	0.338	0.341	93.7
	β_{12}	-0.024	0.579	0.596	95.2	-0.007	0.558	0.598	96.7
	β_{21}	-0.008	0.263	0.257	95.0	0.050	0.262	0.255	93.1
	β_{22}	0.009	0.470	0.476	93.8	0.042	0.456	0.462	93.9
	η	-0.011	0.194	0.200	92.6	0.024	0.208	0.193	92.7
2	β_{11}	-0.007	0.431	0.414	93.8	0.059	0.417	0.424	94.3
	β_{12}	0.046	0.749	0.730	93.2	-0.057	0.639	0.723	96.6
	β_{21}	-0.014	0.246	0.253	95.2	0.062	0.260	0.257	94.5
	β_{22}	0.010	0.475	0.468	94.6	0.040	0.455	0.464	94.1
	η	-0.028	0.183	0.182	90.8	-0.011	0.218	0.187	90.0

Table 4: Simulation results on regression parameters based on the simulated data with different baseline functions.

r	Par.	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
		$(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})=(0.2, 0.2, -0.2, -0.2)$				$(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})=(-0.2, -0.2, -0.2, -0.2)$			
0	β_{11}	-0.016	0.228	0.260	96.0	-0.031	0.264	0.260	92.7
	β_{12}	-0.010	0.439	0.460	94.8	-0.041	0.438	0.453	94.2
	β_{21}	-0.030	0.262	0.276	94.4	-0.044	0.247	0.250	93.6
	β_{22}	-0.061	0.531	0.504	92.0	-0.035	0.452	0.446	93.1
	η	0.042	0.227	0.246	95.2	-0.067	0.167	0.208	91.8
0.5	β_{11}	-0.042	0.290	0.305	95.2	-0.023	0.293	0.298	94.0
	β_{12}	-0.003	0.481	0.530	96.2	-0.041	0.527	0.522	94.9
	β_{21}	-0.038	0.272	0.270	93.8	-0.037	0.242	0.247	94.2
	β_{22}	-0.051	0.471	0.498	93.6	-0.009	0.428	0.453	93.7
	η	0.017	0.228	0.230	95.0	-0.064	0.166	0.196	91.5
1	β_{11}	-0.067	0.326	0.343	94.2	-0.016	0.356	0.333	91.6
	β_{12}	-0.016	0.587	0.600	94.0	-0.021	0.583	0.594	94.1
	β_{21}	-0.049	0.252	0.268	96.0	-0.036	0.250	0.251	92.7
	β_{22}	-0.029	0.504	0.496	93.4	-0.021	0.462	0.449	93.2
	η	0.018	0.220	0.224	94.4	-0.080	0.173	0.196	90.7
2	β_{11}	-0.019	0.440	0.426	93.0	-0.014	0.405	0.401	94.8
	β_{12}	-0.029	0.734	0.737	94.0	-0.028	0.709	0.713	94.8
	β_{21}	-0.043	0.242	0.263	95.6	-0.036	0.236	0.244	94.0
	β_{22}	-0.071	0.486	0.484	95.0	-0.037	0.442	0.456	95.5
	η	-0.003	0.214	0.210	94.0	-0.086	0.177	0.196	89.8

Table 5: Simulation results on regression parameters based on the misspecified distribution with one covariate $X \sim B(1, 0.5)$.

r	Par.	$n = 200$				$n = 400$			
		Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
$(\beta_1, \beta_2) = (0, 0)$									
0	β_1	0.008	0.261	0.260	93.8	0.004	0.186	0.179	92.6
	β_2	-0.011	0.191	0.184	93.8	-0.006	0.130	0.131	95.0
0.5	β_1	-0.012	0.311	0.309	95.0	-0.002	0.200	0.214	95.8
	β_2	0.008	0.185	0.184	95.2	-0.003	0.137	0.130	93.6
1	β_1	-0.002	0.362	0.357	94.4	0.000	0.238	0.244	95.0
	β_2	0.010	0.192	0.183	93.6	-0.008	0.134	0.131	94.4
2	β_1	0.011	0.435	0.435	93.4	-0.009	0.301	0.307	94.6
	β_2	0.003	0.186	0.184	94.0	-0.005	0.135	0.132	94.2
$(\beta_1, \beta_2) = (0.5, 0.5)$									
0	β_1	0.047	0.283	0.272	94.3	0.033	0.192	0.190	92.4
	β_2	0.013	0.192	0.191	94.7	0.010	0.138	0.137	93.7
0.5	β_1	0.032	0.319	0.315	93.9	0.022	0.217	0.219	94.9
	β_2	0.017	0.200	0.189	94.9	0.011	0.138	0.134	93.5
1	β_1	-0.009	0.359	0.356	94.1	0.001	0.242	0.246	94.4
	β_2	0.008	0.191	0.187	96.1	0.009	0.138	0.133	93.0
2	β_1	-0.017	0.423	0.431	94.9	-0.015	0.289	0.303	94.6
	β_2	0.014	0.183	0.186	94.3	0.012	0.129	0.133	95.6

Table 6: Estimated covariate effects for the lung tumor study.

Degree	Model	$\hat{\beta}_1$	SEE	p -value	$\hat{\beta}_2$	SEE	p -value	$\hat{\eta}$	SEE	p -value
With informative censoring										
2	PH frailty	0.599	0.423	0.157	-2.104	0.289	< 0.001	0.080	0.040	0.043
	PO frailty	1.020	0.592	0.085	-2.101	0.201	< 0.001	0.079	0.039	0.044
	Optimal model	0.691	0.429	0.107	-2.119	0.283	< 0.001	0.090	0.034	0.008
3	PH frailty	0.674	0.448	0.132	-2.091	0.289	< 0.001	0.072	0.033	0.031
	PO frailty	1.061	0.590	0.072	-2.100	0.204	< 0.001	0.078	0.039	0.043
	Optimal model	0.769	0.470	0.102	-2.113	0.247	< 0.001	0.086	0.035	0.013
4	PH frailty	0.604	0.362	0.095	-2.129	0.266	< 0.001	0.096	0.037	0.010
	PO frailty	1.010	0.552	0.067	-2.075	0.247	< 0.001	0.062	0.022	0.006
	Optimal model	0.747	0.444	0.092	-2.117	0.228	< 0.001	0.088	0.043	0.041
Without informative censoring										
2	PH frailty	0.933	0.401	0.020	-1.966	0.258	< 0.001	-	-	-
	PO frailty	1.845	0.430	< 0.001	-1.966	0.268	< 0.001	-	-	-
	Optimal model	1.080	0.463	0.020	-1.966	0.273	< 0.001	-	-	-
3	PH frailty	0.933	0.391	0.017	-1.966	0.229	< 0.001	-	-	-
	PO frailty	1.842	0.424	< 0.001	-1.966	0.242	< 0.001	-	-	-
	Optimal model	1.074	0.452	0.017	-1.966	0.275	< 0.001	-	-	-
4	PH frailty	0.937	0.383	0.014	-1.966	0.305	< 0.001	-	-	-
	PO frailty	2.058	0.471	< 0.001	-1.966	0.260	< 0.001	-	-	-
	Optimal model	1.074	0.501	0.032	-1.966	0.275	< 0.001	-	-	-