Published in final edited form as:

Biometrics. 2014 September ; 70(3): 579–587. doi:10.1111/biom.12163.

Accelerated Intensity Frailty Model for Recurrent Events Data

Bo Liu¹, Wenbin Lu¹, and Jiajia Zhang²

Bo Liu: bliu4@ncsu.edu; Wenbin Lu: lu@stat.ncsu.edu; Jiajia Zhang: jzhang@mailbox.sc.edu

¹Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A.

²Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, South Carolina, U.S.A.

Summary

In this article we propose an accelerated intensity frailty (AIF) model for recurrent events data and derive a test for the variance of frailty. In addition, we develop a kernel-smoothing-based EM algorithm for estimating regression coefficients and the baseline intensity function. The variance of the resulting estimator for regression parameters is obtained by a numerical differentiation method. Simulation studies are conducted to evaluate the finite sample performance of the proposed estimator under practical settings and demonstrate the efficiency gain over the Gehan rank estimator based on the AFT model for counting process (Lin et al., 1998, *Biometrika* 85, 605–618). Our method is further illustrated with an application to a bladder tumor recurrence data.

Keywords

Accelerated intensity frailty model; EM Algorithm; Kernel smoothing; Nonparametric maximum likelihood estimation; Recurrent events data

1. Introduction

Recurrent events data are frequently encountered in areas of industry, economics, public health and medical research, where subjects may experience repeated occurrences of the same type of events. Examples include recurrent tumors, infections, asthma attacks and car warranty claims, and many others (Cook and Lawless, 2007). A number of methods have been developed for analysis of recurrent events data. One popular approach is to model the intensity function of the counting process for recurrent events, such as the well known Andersen-Gill model (Andersen and Gill, 1982), which assumes a Cox-type multiplicative intensity function. Under the Andersen-Gill model, the underlying counting process becomes a nonhomogeneous Poisson process adjusting for covariates. However, the multiplicative intensity function is likely to be too stringent in practice due to the within-subject dependency and it generally requires the adjustment of complicated time-dependent covariates. Alternatively, marginal approaches have been widely studied, for example, Cox

Supplementary Materials

The Web Appendix referenced in Sections 4 and 5 are available with this paper at the *Biometrics* website on Wiley Online Library. The bladder tumor recurrence data considered in Section 5 and the associated R codes are also available online.

regression model for the marginal hazard functions of time to recurrent events (Wei et al., 1989) and the marginal models for the mean/rate functions of the counting process for recurrent events (Pepe and Cai, 1993; Lawless and Nadeau, 1995; Lawless et al., 1997; Lin et al., 2000). In many applications, the within-subject dependence of recurrent events is also of primary interest. To study the within-subject dependence, the random effect/frailty models have been proposed, for example the multiplicative intensity frailty model (Andersen et al., 1993). Its associated nonparametric maximum likelihood estimation and asymptotic theory have been investigated by many authors (e.g. Murphy, 1994, 1995; Parner, 1998).

In addition to the Cox-type regression models for recurrent events data, many other semiparametric models have also been studied in the literature, such as the accelerated failure time (AFT) model for counting process (Lin et al., 1998; Jin et al., 2006), the linear transformation models for gap times or counting process (Lu, 2005; Zeng and Lin, 2007) and the additive intensity/rate models (Schaubel et al., 2006; Liu and Wu, 2011). They provide useful alternatives for the Cox-type regression models.

In this paper, we propose an accelerated intensity frailty (AIF) model for recurrent events data. A similar model was also studied by Strawderman (2006) for gap times using an EM-like algorithm. However, Strawderman's estimator does not maximize the likelihood function and its variance needs to be obtained using the bootstrap method. Here, we develop the nonparametric maximum likelihood estimation (NPMLE) method for the proposed model based on a kernel-smoothing aided EM algorithm. The remainder of the paper is organized as follows. Section 2 introduces the AIF model for recurrent events data. A test for the variance of the frailty is derived in Section 3. Section 4 presents our NPMLE estimators and the corresponding variance estimation procedure via numerical differentiation. The large sample properties of the NPMLE estimators are also briefly discussed. Simulation studies and an application to a bladder tumor recurrence data are provided in Section 5. Some conclusions and discussions are given in Section 6.

2. AIF Model

Let $N_i^*(t)$, $i = 1, \dots, n$, be the number of events observed on subject i by time t in the absence of censoring, and $\mathbf{X}_i$ be a p -dimensional vector of baseline covariates. With censoring, we only observe $N_i(t) = N_i^*(t \wedge C_i)$, where C_i is the censoring time and $a \wedge b$ denotes the minimum of a and b . Then $N_i(t)$ is a counting process with the jumps only at the observed recurrence times $0 < T_{i1} < \dots < T_{i, n_i} \leq C_i$, where T_{ij} is the time to the j th recurrence and n_i is the total number of observed recurrences on subject i . Given covariates $\mathbf{X}_i$ and a random effect α_i , the conditional intensity function of $N_i^*(t)$ is specified by

$$\lambda(t|\mathbf{X}_i, \alpha_i) = \alpha_i e^{\beta_0' \mathbf{X}_i} \lambda_0(t e^{\beta_0' \mathbf{X}_i}), \quad (1)$$

where $\lambda_0(\cdot)$ is an unspecified baseline intensity function, and α_i is a positive random variable with mean 1 to ensure the identifiability of the baseline intensity function and is assumed to be independent of $\mathbf{X}_i$. When there is only one recurrent event per subject, model (1) becomes the accelerated failure time frailty model for survival times, which has been studied by

various authors (e.g. Lambert et al., 2004; Klein et al., 1999; Pan, 2001; Zhang and Peng, 2007). Under model (1), conditional on α_i and $\mathbf{X}_i$, $N_i^*(t)$ is a nonhomogeneous Poisson process with mean $\alpha_i \Lambda_0(t e^{\beta_0' \mathbf{X}_i})$, where $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$. Therefore, we have

$E\{N_i^*(t) | \mathbf{X}_i\} = \Lambda_0(t e^{\beta_0' \mathbf{X}_i})$, which coincides with the marginal AFT model for counting process studied by Lin et al. (1998) and Jin et al. (2006). Here the frailty α_i is to describe the association among gap times of consecutive events for the same subject, the larger the variance of α_i the stronger the association. Throughout the paper, it is assumed that C_i is independent of $N_i^*(\cdot)$ given $\{\mathbf{X}_i, \alpha_i\}$ and C_i is independent of α_i .

3. Frailty Variance Test based on Gehan Rank Estimator

It is of interest to test whether the variance of the frailty is zero in the AIF model, i.e. $H_0 : \text{Var}(\alpha_i) = 0$, since it implies that gap times between consecutive recurrences are conditionally independent given covariates. Note that the AIF model implies the marginal AFT model for counting process (Lin et al., 1998). Therefore, the Gehan rank estimator proposed by Lin et al. (1998) is a consistent estimator regardless of the distribution of the frailty including the case with $\alpha_i \equiv 1$. Let $\hat{\beta}_G$ denote the Gehan rank estimator of β_0 and

$\hat{\Lambda}_G(t; \hat{\beta}_G) = \sum_{i=1}^n \int_0^t dN_i(s e^{-\hat{\beta}_G' \mathbf{X}_i}) / \sum_{j=1}^n Y_j(s; \hat{\beta}_G)$ denote the estimator of Λ_0 proposed by Lin et al. (1998), where $Y_j(s; \beta) = I\{C_j(\beta) \geq s\}$ and $C_j(\beta) = C_j e^{\beta' \mathbf{X}_j}$. Under H_0 , the total

number of events n_i on subject i is a Poisson random variable with mean $\Lambda_0(C_i e^{\beta_0' \mathbf{X}_i})$ and variance $\Lambda_0(C_i e^{\beta_0' \mathbf{X}_i})$, while under the alternative $\text{Var}(\alpha_i) > 0$, the variance of n_i is larger than the mean. This motivates us to consider the following test statistic

$$T_n = \sum_{i=1}^n \left[\left\{ n_i - \hat{\Lambda}_G(C_i e^{\hat{\beta}_G' \mathbf{X}_i}; \hat{\beta}_G) \right\}^2 - \hat{\Lambda}_G(C_i e^{\hat{\beta}_G' \mathbf{X}_i}; \hat{\beta}_G) \right],$$

and reject H_0 when $T_n > c$. A similar test statistic is also studied in Dean and Lawless (1989) for detecting overdispersion in Poisson regression. Under H_0 , it can be shown that $n^{-1/2} T_n$ is asymptotically normally distributed with mean zero based on the asymptotic results established in Lin et al. (1998). To compute the variance of the test statistic T_n , we adopt the resampling technique developed by Lin et al. (1998). Specifically, we consider the perturbed test statistic:

$$\begin{aligned} T_n^* = & \sum_{i=1}^n \left[\hat{\Lambda}_G^2(C_i e^{\hat{\beta}_G' \mathbf{X}_i}; \hat{\beta}_G) - \hat{\Lambda}_G^2(C_i e^{\hat{\beta}_G^{*'} \mathbf{X}_i}; \hat{\beta}_G) - (2n_i + 1) \left\{ \hat{\Lambda}_G(C_i e^{\hat{\beta}_G' \mathbf{X}_i}; \hat{\beta}_G) - \hat{\Lambda}_G(C_i e^{\hat{\beta}_G^{*'} \mathbf{X}_i}; \hat{\beta}_G) \right\} \right] \\ & + \sum_{i=1}^n (V_i - 1) \left\{ n_i^2 + \hat{\Lambda}_G^2(C_i e^{\hat{\beta}_G' \mathbf{X}_i}; \hat{\beta}_G) - (2n_i + 1) \hat{\Lambda}_G(C_i e^{\hat{\beta}_G' \mathbf{X}_i}; \hat{\beta}_G) \right\} \\ & + \sum_{i=1}^n \hat{W}_i \left\{ 2\hat{\Lambda}_G(C_i e^{\hat{\beta}_G' \mathbf{X}_i}; \hat{\beta}_G) - (2n_i + 1) \right\}, \end{aligned}$$

where $V_1, \dots, V_n$ are i.i.d. positive random variables with mean 1 and variance 1,

$$\hat{W}_i = \hat{\Lambda}_G(C_i e^{\hat{\beta}_G' \mathbf{X}_i}; \hat{\beta}_G) - \hat{\Lambda}_G(C_i e^{\hat{\beta}_G^* \mathbf{X}_i}; \hat{\beta}_G^*) + \sum_{j=1}^n (V_j - 1) \int_0^{C_i e^{\hat{\beta}_G^* \mathbf{X}_i}} \frac{dM_j(s; \hat{\beta}_G)}{\sum_{k=1}^n Y_k(s; \hat{\beta}_G)},$$

$dM_j(t; \beta) = dN_j(t e^{-\beta' \mathbf{X}_j}) - Y_j(t; \beta) d\Lambda_G(t; \beta)$, $\hat{\beta}_G^*$ is the perturbed estimator for β obtained based on the same set of random variables V_i 's as proposed in Lin et al. (1998). Following the same arguments as in Lin et al. (1998), it can be shown that conditional on observed data, $n^{-1/2} T_n^*$ has the same limiting distribution as $n^{-1/2} T_n$ under H_0 . Therefore, we can use the empirical distribution of $n^{-1/2} T_n^*$ to calculate the critical value c .

4. Proposed Estimation Method

Assume that $\text{Var}(a_i) > 0$. We develop an EM algorithm to obtain the nonparametric maximum likelihood estimators for the AIF model. Specifically, let $f_\alpha(\cdot; \theta)$ denote the density function of a_i , where θ is a scalar parameter related to the variance of frailty. In practice, a variety of distributions can be used for the frailty, such as gamma distribution (Clayton, 1978) and log-normal distribution (McGilchrist and Aisbett, 1991). Then the complete log-likelihood, up to a constant independent of parameters, can be written as

$$l_n^c(\beta, \Lambda, \theta) = l_{n,1}^c(\theta) + l_{n,2}^c(\beta, \Lambda),$$

where

$$l_{n,1}^c(\theta) = n^{-1} \sum_{i=1}^n \{n_i \log \alpha_i + \log f_\alpha(\alpha_i; \theta)\}, \quad (2)$$

$$l_{n,2}^c(\beta, \Lambda) = n^{-1} \sum_{i=1}^n \left(\sum_{j=1}^{n_i} [\beta' \mathbf{X}_i + \log \lambda \{ \tilde{T}_{ij}(\beta) \}] - \alpha_i \Lambda \{ \tilde{C}_i(\beta) \} \right), \quad (3)$$

with $\Lambda(t) = \int_0^t \lambda(s) ds$ and $\tilde{T}_{ij}(\beta) = T_{ij} e^{\beta' \mathbf{X}_i}$.

In the E-step, we compute the conditional expectations $E(a_i | \mathcal{O}, \Omega^{[k]})$, $E(\log a_i | \mathcal{O}, \Omega^{[k]})$ and $E\{\log f_\alpha(a_i; \theta) | \mathcal{O}, \Omega^{[k]}\}$, where $\Omega^{[k]} = (\hat{\beta}^{[k]}, \hat{\Lambda}^{[k]}, \hat{\theta}^{[k]})$ denote the current estimates at step k and $\mathcal{O} = \{(T_{ij}, C_i, \mathbf{X}_i), i = 1, \dots, n; j = 1, \dots, n_i\}$ denote the observed data. The above expectations can be calculated as the integrals of the corresponding terms with respect to the following conditional density:

$$f(a_i | \mathcal{O}, \hat{\Omega}^{[k]}) = \frac{f_\alpha(a_i; \hat{\theta}^{[k]}) \alpha_i^{n_i} e^{-\alpha_i \hat{\Lambda}^{[k]} \{ \tilde{C}_i(\hat{\beta}^{[k]}) \}}}{\int_0^\infty f_\alpha(a_i; \hat{\theta}^{[k]}) \alpha_i^{n_i} e^{-\alpha_i \hat{\Lambda}^{[k]} \{ \tilde{C}_i(\hat{\beta}^{[k]}) \}} d\alpha_i}.$$

For instance, when the frailty follows a gamma distribution with the density $f_{\alpha}(x; \theta) = x^{\theta-1}e^{-\theta x}\beta/\Gamma(\theta)$ with $\Gamma(\theta) = \int_0^{\infty} t^{\theta-1}e^{-t}dt$, we have

$$\begin{aligned}\hat{\alpha}_i^{[k]} &\equiv E(\alpha_i | \mathcal{O}, \hat{\Omega}^{[k]}) \\ &= (n_i + \hat{\theta}^{[k]}) / [\hat{\theta}^{[k]} \\ &\quad + \hat{\Lambda}^{[k]} \left\{ \tilde{C}_i(\hat{\beta}^{[k]}) \right\}], E_{2,i}^{[k]} \equiv E(\log \alpha_i | \mathcal{O}, \hat{\Omega}^{[k]}) = \Psi(n_i + \hat{\theta}^{[k]}) - \log[\hat{\theta}^{[k]} + \hat{\Lambda}^{[k]} \left\{ \tilde{C}_i(\hat{\beta}^{[k]}) \right\}],\end{aligned}$$

and

$$E_{3,i}^{[k]} \equiv E \left\{ \log f_{\alpha}(\alpha_i; \theta) | \mathcal{O}, \hat{\Omega}^{[k]} \right\} = (\theta - 1)E_{2,i}^{[k]} - \theta \hat{\alpha}_i^{[k]} + \theta \log \theta - \log \Gamma(\theta),$$

where $\Psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. For other frailty distributions where there are no closed forms for these conditional expectations, such as the log-normal distribution, we use gaussian quadrature method to calculate the integrals. Therefore, the conditional expectations of (2) and (3) given $\mathcal{O}$ and $\Omega^{[k]}$ are respectively,

$$E \left\{ l_{n,2}^c(\theta) | \mathcal{O}, \hat{\Omega}^{[k]} \right\} = n^{-1} \sum_{i=1}^n (n_i E_{2,i}^{[k]} + E_{3,i}^{[k]}), \quad (4)$$

$$E \left\{ l_{n,2}^c(\beta, \Lambda) | \mathcal{O}, \hat{\Omega}^{[k]} \right\} = n^{-1} \sum_{i=1}^n \left(\sum_{j=1}^{n_i} [\beta' \mathbf{X}_i + \log \lambda \left\{ \tilde{T}_{ij}(\beta) \right\}] - \hat{\alpha}_i^{[k]} \Lambda \left\{ \tilde{C}_i(\beta) \right\} \right). \quad (5)$$

In the M-step, we maximize (4) and (5), respectively. Maximization of (4) can be achieved easily using standard gradient-based algorithms. Here, we used the “optim” function in R, and denote the maximizer of (4) by $\theta^{[k+1]}$. Based on model (1) and the conditional independent censoring assumption, the intensity function of the counting process

$N_i(t; e^{-\beta_0' \mathbf{X}_i})$ given α_i and $\mathbf{X}_i$ is $I\{C_i(\beta) \geq t\} \alpha_i \lambda_0(t)$. Therefore, given $\beta_0 = \beta$ and $\alpha_i = \hat{\alpha}_i^{[k]}$, the nonparametric maximum likelihood estimator of $\Lambda_0(t)$ is given by

$$\hat{\Lambda}^{[k]}(t; \beta) = \sum_{i=1}^n \int_0^t \frac{dN_i(se^{-\beta' \mathbf{X}_i})}{\sum_{j=1}^n I\left\{ \tilde{C}_j(\beta) \geq s \right\} \hat{\alpha}_j^{[k]}}.$$

Let $\Lambda^{[k]}(t; \beta)$ denote the jump of $\hat{\Lambda}^{[k]}(t; \beta)$ at t . Due to the fact that

$\sum_{i=1}^n \hat{\alpha}_i^{[k]} \hat{\Lambda}^{[k]} \left\{ \tilde{C}_i(\beta); \beta \right\} = \sum_{i=1}^n n_i$, the resulting profile likelihood function of β from (5) is

$$\tilde{l}_{n,2}^{[k]}(\beta) = n^{-1} \sum_{i=1}^n \left(\sum_{j=1}^{n_i} [\beta' \mathbf{X}_i + \log \Delta \hat{\Lambda}^{[k]} \{ \tilde{T}_{ij}(\beta); \beta \}] - n_i \right),$$

which is discrete and can not achieve its maximum with finite values of β as noted by Zeng and Lin (2007) for the nonparametric maximum likelihood estimation in the standard AFT model. Smoothing is needed to approximate maximization. Note that

$$d\hat{\Lambda}^{[k]}(t; \beta) = \sum_{i=1}^n dN_i(t e^{-\beta' \mathbf{X}_i}) / \sum_{j=1}^n I \{ \tilde{C}_j(\beta) \geq t \} \hat{\alpha}_j^{[k]} \text{ and the indicator function}$$

$I \{ \tilde{C}_j(\beta) \geq t \}$ can be approximated by the smooth function $\int_{-\infty}^{\{\tilde{C}_j(\beta) - t\}/h_n} K(u) du$ as $h_n \rightarrow 0$, where $K(\cdot)$ is a smooth symmetric kernel density with bandwidth h_n . Then, given β and

$\hat{\alpha}_i^{[k]}$, a smoothed nonparametric maximum likelihood estimator of $\lambda_0(t)$ can be obtained by

$$\hat{\lambda}^{[k]}(t; \beta) = \frac{h_n^{-1} \sum_{i=1}^n \int_0^\infty K \{ (s - t)/h_n \} dN_i(s e^{-\beta' \mathbf{X}_i})}{\sum_{j=1}^n \hat{\alpha}_j^{[k]} \int_{-\infty}^{\{\tilde{C}_j(\beta) - t\}/h_n} K(u) du} - \frac{h_n^{-1} \sum_{i=1}^n \sum_{j=1}^{n_i} K \{ \{ \tilde{T}_{ij}(\beta) - t \} / h_n \}}{\sum_{j=1}^n \hat{\alpha}_j^{[k]} \int_{-\infty}^{\{\tilde{C}_j(\beta) - t\}/h_n} K(u) du}.$$

Therefore, the smoothed profile likelihood function of β is

$$\tilde{l}_{n,2}^{[k]}(\beta) = n^{-1} \sum_{i=1}^n \left(\sum_{j=1}^{n_i} [\beta' \mathbf{X}_i + \log \hat{\lambda}^{[k]} \{ \tilde{T}_{ij}(\beta); \beta \}] - n_i \right). \quad (6)$$

Denote the maximizer of (6) by $\hat{\Lambda}^{[k+1]}(t) = \int_0^t \hat{\lambda}^{[k]}(s; \hat{\beta}^{[k+1]}) ds$.

From initial estimators $\Omega^{[0]} = (\hat{\beta}^{[0]}, \hat{\theta}^{[0]}, \hat{\Lambda}^{[0]})$, we iterate between the E-step and M-step until a pre-determined convergence criterion is satisfied. In our implementation, we chose $\hat{\beta}^{[0]} = \hat{\beta}_G$, $\hat{\theta}^{[0]} = 1$. Setting $k = -1$ and $\hat{\alpha}_i^{[-1]} \equiv 1$ in $\hat{\lambda}^{[k]}(t; \hat{\beta}^{[0]})$, we obtain $\hat{\Lambda}^{[0]}$. Our limited numerical experience suggests that the proposed EM algorithm is not sensitive to the choice of $\theta^{[0]}$ since $\hat{\beta}_G$ is a consistent estimator of β_0 regardless of the frailty distribution. Denote the estimators of β , θ and $\Lambda(\cdot)$ at convergence by $\hat{\beta}_n$, $\hat{\theta}_n$ and $\hat{\Lambda}_n(\cdot)$, respectively. Following the techniques used by Liu et al. (2013) for studying the kernel-smoothing aided nonparametric maximum likelihood estimation in the AFT frailty model with clustered survival data, it can be shown that under certain regularity conditions, and $nh_n^2 \rightarrow \infty$ and $nh_n^4 \rightarrow 0$ as $n \rightarrow \infty$, the proposed estimators are consistent, and $\hat{\beta}_n$ and $\hat{\theta}_n$ are asymptotically normal and semiparametrically efficient. The proofs are given in the Web Appendix.

To estimate the variance of $\hat{\beta}_n$, we use the EM-aided numerical differentiation method proposed by Chen and Little (1999), which numerically computes the empirical Fisher Information matrix of the observed profile likelihood. Specifically, write

$$E \left\{ l_{n,1}^c(\theta) + l_{n,2}^c(\beta, \Lambda) \mid \mathcal{O}, \Omega \right\} = \frac{1}{n} \sum_{i=1}^n \tilde{l}_i(\beta, \Lambda, \theta).$$

We perturb the j th component of $\hat{\beta}_n$ by a small value d and denote the pair of perturbed estimates by $\hat{\beta}_{n,j-} = (\hat{\beta}_{n,1}, \dots, \hat{\beta}_{n,j} - d, \dots, \hat{\beta}_{n,p})$ and $\hat{\beta}_{n,j+} = (\hat{\beta}_{n,1}, \dots, \hat{\beta}_{n,j} + d, \dots, \hat{\beta}_{n,p})$ for $j = 1, \dots, p$. Fix β at $\hat{\beta}_{n,j-}$ and run the EM Algorithm until convergence. Denote the estimates of Λ and θ at convergence by $\hat{\Lambda}_{n,j-}$ and $\hat{\theta}_{n,j-}$, respectively. Similarly, we can obtain the estimates $\hat{\Lambda}_{n,j+}$ and $\hat{\theta}_{n,j+}$. For $i = 1, \dots, n$ and $j = 1, \dots, p$, define

$$\tilde{S}_{ij} = \left\{ \tilde{l}_i(\hat{\beta}_{n,j+}, \hat{\Lambda}_{n,j+}, \hat{\theta}_{n,j+}) - \tilde{l}_i(\hat{\beta}_{n,j-}, \hat{\Lambda}_{n,j-}, \hat{\theta}_{n,j-}) \right\} / (2d).$$

Let $\tilde{\mathbf{S}}_i = (S_{i1}, \dots, S_{ip})'$ and $\hat{I}_n = \sum_{i=1}^n \tilde{\mathbf{S}}_i \tilde{\mathbf{S}}_i'$. Then the variance-covariance matrix of $\hat{\beta}_n$ can be estimated by $\hat{I}_n^{-1}$.

5. Numerical Studies

5.1 Simulations

We generate recurrent events data from the proposed AIF model with two covariates $\mathbf{X}_i = (X_{i1}, X_{i2})'$, where X_{i1} follows Bernoulli(0.5) and X_{i2} follows uniform $[-1, 1]$. In addition, we consider two frailty distributions: gamma frailty with mean 1 and variance $\sigma_2 = 1/\theta$ and log-normal frailty with mean 1 and variance $\sigma_2 = e^\theta - 1$; and two functions for $\Lambda_0(t)$: $\Lambda_0(t) = \log(t + 1)$ and $\Lambda_0(t) = -\log[1 - \Phi\{\log(t)\}]$, where Φ is the standard normal cumulative distribution function. We set $\beta_0 = (-1, 1)'$ and $\sigma_2 = 2, 1$ or 0.5 for gamma frailty and $2.32, 1.72$ or 0.65 for log-normal frailty. The censoring time C_i 's are generated from uniform $[0, \tau_c]$, where τ_c is chosen to yield an average of 2 and 4 events per subject. Given a_i , C_i and $\mathbf{X}_i$, the number of observed recurrent events n_i for subject i is generated from a poisson

distribution with mean $\alpha_i \Lambda_0(C_i e^{\beta_0' \mathbf{X}_i})$. Following similar arguments given in Liang et al. (2009), conditional on $(C_i, \mathbf{X}_i, n_i)$, the recurrent event times $0 < T_{i1} < \dots < T_{in_i} < C_i$ of subject i are obtained as the order statistics of a set of i.i.d random variables with the joint density function

$$p(t_{i1}, \dots, t_{in_i} \mid C_i, \mathbf{X}_i, n_i) = n_i! \prod_{j=1}^{n_i} \frac{\lambda_0(t_{ij} e^{\beta_0' \mathbf{X}_i}) e^{\beta_0' \mathbf{X}_i}}{\Lambda_0(C_i e^{\beta_0' \mathbf{X}_i})}.$$

For each scenario, we consider 500 simulation runs with the sample size of $n = 100$. In our method, we use Gaussian kernel for computational convenience. But other kernel functions, such as Epanechnikov kernel, can also be used. To select the bandwidth parameter h_n , we set $h_n = \zeta n^{-1/3}$ based on the asymptotic results, where ζ is a positive constant. In our simulation study, we find that ζ between 2 and 3 generally works well in all scenarios for the gamma and log-normal frailties. In general, a cross-validation method can be used to select the best ζ .

For comparison, we also considered the Gehan rank estimator (denoted by GehanR) proposed by Lin et al. (1998). The simulation results are summarized in Tables 1–2. We observe that both estimators for the regression parameters are nearly unbiased under all settings and the averages of estimated standard errors obtained using the proposed EM-aided numerical differentiation method for the NPMLE are close to their standard deviations with the empirical coverage probabilities of 95% confidence intervals close to the nominal level. In addition, the NPMLE is more efficient than the GehanR in almost all cases, and the gain is more substantial when the variance of frailty is big, for example, when $\sigma^2=2$ for gamma frailty and $\Lambda_0(t) = \log(t+1)$, the relative efficiency (RE) for β_1 (defined as the sample variance of GehanR divided by the sample variance of NPMLE) is 1.88 and 5.59 for 2 and 4 events per subject, respectively, and it is 2.21 and 7.43 for log-normal frailty with $\sigma_2 = 2.32$. This agrees with our expectation since when there is a strong dependence among recurrent event times, by effectively accounting for the association via a frailty, the NPMLE is expected to be more efficient than the GehanR assuming the working independence. Moreover, the biases of the NPMLE for the variance of the frailty are relatively small, especially when the number of events per subject is large. We note that the biases of the frailty variance estimates under the log-normal frailty are relatively larger than those under the gamma frailty, which may be partly due to the numerical integration used in the kernel smoothing based EM algorithm. To assess the estimation of $\Lambda_0(t)$ using our method, we plot the mean of estimated curves $\hat{\Lambda}_n(t)$ versus the true curve $\Lambda_0(t)$, see the first and second rows of Figure 1. For illustration, here we only report the results for the cases with 4 averaged number of events per subject and $\Lambda_0(t) = \log(t+1)$. It is observed that the estimated curves are close to the true curves for all the settings.

Next, we conduct a sensitivity analysis to study the performance of the proposed NPMLE when the frailty distribution is misspecified. Specifically, here the frailty is generated from the log-normal distribution, but the NPMLE is computed based on the gamma frailty in the EM algorithm. The simulation results are given in Table 3. It can be seen that the NPMLE for the regression parameters still shows very small biases that are comparable to those reported in Table 2 when the log-normal frailty distribution is correctly specified, and the means of estimated standard errors are close to the standard deviations with proper coverage probabilities. However, the variance of the frailty is seriously underestimated when the log-normal frailty is misspecified as the gamma frailty. In addition, based on the NPMLE estimates of $\Lambda_0(t)$ given in the third row of Figure 1, the mean estimated curves are still close to the true curves under the misspecification of the frailty distribution. In summary, the NPMLE shows relatively robust performance to the misspecification of the frailty distribution for estimation of β_0 and $\Lambda_0(\cdot)$. One possible explanation is that parameters β_0 and $\Lambda_0(\cdot)$ in the AIF model have the same marginal interpretation as those in the AFT model for counting process (Lin et al., 1998).

In addition, we compare our proposed NPMLE with a maximum kernel smoothed pseudo likelihood estimator assuming working independence, denoted by MPSLE. Specifically, to obtain MPSLE, we set $\alpha_i \equiv 1$, $i = 1, \dots, n$, in the log likelihood function $l_{n,2}^c(\beta, \Lambda)$ given in (3), and apply similar kernel smoothing techniques to obtain the smoothed pseudo profile log likelihood function for β . MPSLE of β is the maximizer of the resulting smoothed

pseudo profile log likelihood function. We consider the same simulation settings with $\Lambda_0(t) = \log(t + 1)$. The results are given in the Supplementary Materials. Based on the results, we observe that our proposed NPMLE is more efficient than MPSLE for almost all the scenarios, and the efficiency gain is big for the cases with large frailty variance and large average number of recurrences per subject. In addition, MPSLE is generally more efficient than Gehan rank estimator. Finally, we conduct the proposed frailty variance test in Section 3 and consider the same simulation settings with the average number of recurrence per subject as 2. Overall, the proposed test gives reasonable power for both gamma and log-normal frailties, and two choices of baseline intensity functions. For example, for gamma frailty, the power ranges from 0.835 to 0.945, while for the log-normal frailty, it ranges from 0.635 to 0.77. The detailed simulation results are given in the Supplementary Materials.

5.2 Analysis of Bladder Tumor Recurrence Data

We apply our estimation method to a bladder tumor recurrence data from a study conducted by the Veterans Administration Co-operative Urological Group between 1971 and 1976. The study enrolled 118 patients with bladder tumor, who were randomly assigned to one of the three treatment groups, placebo, thiotepa and pyridoxine. As in the analysis conducted by Lin et al. (1998), we only focus on two groups, placebo and thiotepa, with the total number of patients $n = 86$. Among them, 48 received placebo with 87 total recurrences, while 38 received thiotepa with 45 total recurrences. The maximum number of recurrences is 9, but there are 39 patients who never experienced any tumor recurrence in the study duration, showing the big heterogeneity among patients in terms of tumor recurrence, partly due to the differences in covariates and censoring times. The prognostic covariates include treatment (1 if treated with placebo and 0 otherwise), diameter of the largest initial tumor (in centimeters), and number of initial tumors.

Lin et al. (1998) conducted a goodness-of-fit test for the counting process based marginal AFT model, and found no evidence against this model for the bladder tumor recurrence data. To test the dependence of tumor recurrence times after adjusting for baseline covariates, we conduct the test proposed in Section 3 for the null hypothesis $H_0 : \text{Var}(\alpha_i) = 0$. We consider a one sided test with the significance level 0.05 based on 1000 resampled test statistics. The empirical p-value is 0.022, which rejects the null hypothesis and indicates the within-subject dependency of tumor recurrence times even after adjusting for covariates.

Next, we analyze the data using our proposed method with the gamma and log-normal frailties. The corresponding estimators are denoted by NPMLE_g and NPMLE_l , respectively. The bandwidth parameters are chosen as the same as in the simulation study. For comparison, we also compute the Gehan rank estimator, denoted by GehanR. The results are given in Table 4. We observe that the NPMLE_g and NPMLE_l give very similar estimates, which suggests that the proposed NPMLE method for the regression coefficients is not sensitive to the choice of the frailty distribution. The estimated variance of frailty is 0.837 for the gamma frailty and 0.893 for the log-normal frailty. In addition, our NPMLEs give more significant p-values compared to the GehanR, indicating possible efficiency gain by taking into account the association of tumor recurrence times. Both methods find that treatment and number of initial tumors have significant effects on the mean number of

bladder tumor recurrences, but initial size does not. In particular, the treatment thiotepa is effective in reducing the mean number of bladder tumor recurrences.

6. Concluding Remarks and Discussions

In this paper, we propose an AIF model for recurrent events data and develop its associated nonparametric maximum likelihood estimation using kernel smoothing based EM algorithm. The numerical studies show that our estimators may have substantial efficiency gain compared to the Gehan rank estimator, especially when the dependence of recurrent event times is strong. Moreover, our limited sensitivity analysis shows that the proposed estimators for regression coefficients and baseline cumulative intensity function are not sensitive to the choice of the frailty distribution. Theoretical properties of the proposed estimators under the misspecification of the frailty distribution need to be further investigated along the line of Kosorok et al. (2004).

In the proposed AIF model, the baseline intensity function is multiplied by a frailty variate to facilitate the developed kernel smoothing based EM estimation. It is also of interest to consider the case that the frailty variate affects the time scale directly. The derivation of the associated nonparametric maximum likelihood estimation will become more challenging and deserve a full investigation. In addition, the frailty distribution in the AIF model needs to be specified in advance. Knowing the frailty distribution can help with the prediction, for example, predicting the next recurrent event time given all the previous recurrent event times and covariates. However, there are relatively fewer works available for determining the most appropriate frailty distribution. Among them available, Choi et al. (2001) proposed an empirical Bayes test for checking the gamma frailty assumption in the multiplicative intensity frailty model. It is an interesting question that warrants our future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the associate editor and two referees for their comments that substantially improved the presentation of the article. W. Lu's work was partially supported by NIH/NCI grants R01 CA140632 and P01 CA142538.

References

- Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. Statistical models based on counting processes. New York: Springer-Verlag; 1993.
- Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study. *Annals of Statistics*. 1982; 10:1100–1120.
- Chen HY, Little RJA. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 1999; 94:896–908.
- Choi S, Jin Z, Ying Z. Goodness-of-fit tests for semiparametric models with multiple event-time data. *Statistica Sinica*. 2001; 11:723–736.
- Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*. 1978; 65:141.

- Cook, RJ.; Lawless, JF. Statistics for Biology and Health. Springer; 2007. The statistical analysis of recurrent events. ISBN 978-0-387-69809-0.
- Dean C, Lawless JF. Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association*. 1989; 84:467–472.
- Jin Z, Lin DY, Ying Z. Rank regression analysis of multivariate failure time data based on marginal linear models. *Scandinavian Journal of Statistics*. 2006; 33:1–23.
- Klein JP, Pelz C, Zhang M-j. Modeling random effects for censored data by a multivariate normal regression model. *Biometrics*. 1999; 55:497–506. [PubMed: 11318206]
- Kosorok MR, Lee BL, Fine JP. Robust inference for univariate proportional hazards frailty regression models. *Annals of Statistics*. 2004; 32:1448–1491.
- Lambert P, Collett D, Kimber A, Johnson R. Parametric accelerated failure time models with random effect and an application to kidney transplant survival. *Statistics in Medicine*. 2004; 2004:3177–3192. [PubMed: 15449337]
- Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. *Technometrics*. 1995; 37:158–168.
- Lawless JF, Nadeau C, Cook RJ. Analysis of mean and rate functions for recurrent events. *Proceedings of the 1st Seattle Symposium in Biostatistics: Survival Analysis*. 1997:37–50.
- Liang Y, Lu W, Ying ZL. Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics*. 2009; 65:377–384. [PubMed: 18759841]
- Lin DY, Wei LJ, Yang I, Ying ZL. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B*. 2000; 62:711–730.
- Lin DY, Wei LJ, Ying ZL. Accelerated failure time models for counting processes. *Biometrika*. 1998; 85:605–618.
- Liu B, Lu W, Zhang J. Kernel smoothed profile likelihood estimation in the accelerated failure time frailty model for clustered survival data, accepted by. *Biometrika*. 2013
- Liu Y, Wu Y. Semiparametric additive intensity model with frailty for recurrent events. *Acta Mathematica Sinica, English Series*. 2011; 27:1831–1842.
- Lu W. Marginal regression of multivariate event times based on linear transformation models. *Lifetime Data Analysis*. 2005; 11:389–404. [PubMed: 16133886]
- McGilchrist CA, Aisbett CW. Regression with frailty in survival analysis. *Biometrics*. 1991; 47:461–466. [PubMed: 1912255]
- Murphy SA. Consistency in a proportional hazard model incorporating a random effect. *Annals of Statistics*. 1994; 22:712–731.
- Murphy SA. Asymptotic theory for the frailty model. *Annals of Statistics*. 1995; 23:182–198.
- Pan W. Using frailties in the accelerated failure time model. *Lifetime Data Analysis*. 2001; 7:55–64. [PubMed: 11280848]
- Parner E. Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*. 1998; 26:183–214.
- Pepe MS, Cai J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association*. 1993; 88:811–820.
- Schaubel DE, Zeng D, Cai J. A semiparametric additive rates model for recurrent event data. *Lifetime Data Analysis*. 2006; 12:389–406. [PubMed: 17031499]
- Strawderman R. A regression model for dependent gap times. *The International Journal of Biostatistics*. 2006; 2:1–33.
- Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*. 1989; 84:1065–1073.
- Zeng D, Lin DY. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*. 2007; 102:1387–1396.
- Zeng D, Lin DY. Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association*. 2007; 102:167–180.
- Zhang J, Peng Y. An alternative estimation method for the accelerated failure time frailty model. *Computational Statistics and Data Analysis*. 2007; 51:4413–4423.

Figure 1.

Estimation of $\Lambda_0(t)$: the first row is for the gamma frailty, the second row is for the log-normal frailty and the third row is for the misspecified frailty.

Table 1

Simulation results for gamma frailty: AVG, average number of events per subject; SD, sample standard deviation; SE, mean of estimated standard errors; CP, empirical coverage probability of 95% Wald-type confidence interval; RE, relative efficiency of NPMLE compared with GehanR.

AVG	σ^2	Bias	SD	β	NPMLE			GehanR			
					Bias	SD	SE	CP(%)	Bias	SD	RE
$\Lambda_0(t)=\log(t+1)$											
2	2	-0.117	.454	-1	0.096	.502	.482	92.6	0.087	.689	1.884
				1	-0.063	.435	.427	95.4	-0.026	.582	1.790
4	2	-0.056	.358	-1	0.015	.437	.435	93.8	0.137	1.033	5.588
				1	0.008	.388	.388	96.2	-0.070	.907	5.465
2	1	0.006	.217	-1	0.042	.465	.463	96.0	0.001	.514	1.222
				1	-0.036	.392	.414	94.6	-0.025	.450	1.318
4	1	0.002	.191	-1	-0.004	.410	.419	94.6	0.011	.766	3.491
				1	-0.007	.351	.376	96.2	-0.031	.685	3.809
2	0.5	0.087	.150	-1	0.028	.413	.435	94.6	0.017	.454	1.208
				1	-0.025	.359	.386	96.0	-0.031	.380	1.120
4	0.5	0.022	.107	-1	0.023	.383	.401	95.6	0.040	.649	2.871
				1	-0.016	.351	.358	94.8	-0.064	.547	2.429
$\Lambda_0(t)=-\log\{1-\Phi(\log t)\}$											
2	2	-0.031	.495	-1	0.085	.321	.336	95.8	0.025	.390	1.476
				1	-0.090	.293	.301	95.0	-0.011	.343	1.370
4	2	-0.025	.399	-1	0.066	.365	.402	95.4	0.049	.470	1.658
				1	-0.038	.330	.349	94.8	-0.007	.405	1.506
2	1	0.051	.242	-1	0.024	.279	.305	96.0	-0.002	.308	1.219
				1	-0.031	.246	.276	96.0	-0.007	.276	1.259
4	1	0.013	.201	-1	0.057	.263	.268	95.2	0.007	.336	1.632
				1	-0.037	.237	.242	94.2	-0.009	.304	1.645
2	0.5	0.093	.158	-1	-0.006	.253	.275	96.2	-0.000	.268	1.122
				1	-0.004	.237	.248	95.6	-0.016	.227	0.917
4	0.5	0.023	.117	-1	-0.003	.248	.241	94.6	-0.012	.293	1.396
				1	-0.011	.209	.215	94.2	-0.010	.240	1.319

Table 2

Simulation results for lognormal frailty: AVG, average number of events per subject; SD, sample standard deviation; SE, mean of estimated standard errors; CP, empirical coverage probability of 95% Wald-type confidence interval; RE, relative efficiency of NPMLE compared with GehanR.

AVG	σ^2	Bias	SD	β	NPMLE				GehanR		
					Bias	SD	SE	CP(%)	Bias	SD	RE
$\Lambda_0(t)=\log(t+1)$											
2	2.320	-0.308	.769	-1	0.052	.491	.469	93.4	-0.001	.730	2.210
				1	-0.027	.414	.417	94.6	-0.029	.587	2.010
4	2.320	-0.136	.779	-1	0.046	.398	.460	96.4	0.019	1.085	7.432
				1	-0.022	.367	.411	95.8	-0.025	.876	5.697
2	1.718	-0.121	.572	-1	0.055	.475	.464	94.0	0.003	.663	1.948
				1	0.04	.406	.415	95.2	-0.022	.556	1.875
4	1.718	-0.045	.523	-1	0.016	.406	.410	94.8	0.015	.985	5.886
				1	-0.014	.382	.363	94.0	-0.011	.773	4.095
2	0.649	0.128	.235	-1	0.009	.405	.439	95.4	-0.022	.479	1.399
				1	-0.000	.374	.390	95.8	0.009	.414	1.225
4	0.649	0.061	.189	-1	-0.006	.378	.392	95.8	0.001	.692	3.351
				1	0.028	.344	.346	94.8	0.012	.575	2.794
$\Lambda_0(t)=-\log\{1-\Phi(\log t)\}$											
2	2.320	-0.371	.731	-1	0.045	.329	.309	96.0	-0.015	.413	1.576
				1	-0.054	.287	.279	93.2	-0.011	.359	1.565
4	2.320	-0.305	.744	-1	0.052	.337	.337	95.0	-0.009	.506	2.254
				1	-0.018	.299	.297	94.2	-0.015	.413	1.908
2	1.718	-0.153	.572	-1	0.009	.293	.282	93.8	-0.019	.377	1.656
				1	-0.018	.262	.254	94.4	-0.009	.328	1.567
4	1.718	-0.216	.484	-1	0.052	.286	.292	95.0	-0.009	.451	2.487
				1	-0.05	.257	.257	93.6	-0.012	.370	2.073
2	0.649	0.088	.257	-1	-0.009	.281	.272	94.0	-0.02	.296	1.110
				1	0.026	.233	.245	96.4	0.005	.244	1.097
4	0.649	-0.011	.177	-1	0.006	.243	.249	95.0	-0.017	.322	1.756
				1	-0.012	.215	.222	95.6	0.008	.265	1.519

Table 3

Sensitivity analysis results for misspecified gamma frailty distribution: AVG, average number of events per subject; SD, sample standard deviation; SE, mean of estimated standard errors; CP, empirical coverage probability of 95% Wald-type confidence interval.

AVG	σ^2	$\Lambda_0(t) = \log(t + 1)$						$\Lambda_0(t) = -\log\{1 - \Phi(\log t)\}$					
		Bias(σ^2)	SD(σ^2)	Bias	SD	SE	CP(%)	Bias(σ^2)	SD(σ^2)	Bias	SD	SE	CP(%)
2	2.320	-1.021	.367	0.030	.485	.495	94.6	-0.973	.383	0.019	.313	.331	96.6
				-0.019	.422	.446	95.0			-0.044	.293	.303	94.4
4	2.320	-1.085	.278	0.021	.427	.426	94.8	-1.082	.296	0.033	.322	.348	95.4
				0.005	.392	.384	93.0			-0.017	.285	.308	95.0
2	1.718	-0.632	.294	0.042	.488	.491	93.8	-0.596	.311	0.002	.303	.303	95.2
				0.011	.414	.442	96.8			-0.013	.270	.276	96.2
4	1.718	-0.692	.225	0.014	.419	.420	94.8	-0.678	.242	0.049	.297	.284	94.2
				-0.018	.384	.374	94.2			-0.049	.259	.256	94.2
2	0.649	-0.057	.159	-0.000	.415	.425	95.0	-0.054	.183	-0.013	.286	.271	93.8
				0.012	.388	.381	94.0			0.032	.235	.246	94.6
4	0.649	-0.117	.118	-0.000	.381	.398	96.2	-0.114	.128	0.009	.247	.263	96.4
				0.028	.346	.354	94.8			-0.012	.216	.235	96.2

Table 4

Analysis results for bladder tumor data: Est., point estimates of the regression coefficients; SE, estimated standard errors; pv, p-value of the statistic.

	NPMLE _g			NPMLE _t			GehanR		
	Est.	SE	pv	Est.	SE	pv	Est.	SE	pv
Treatment	0.623	0.274	0.023	0.612	0.244	0.012	0.657	0.314	0.036
Initial number	0.462	0.098	0.000	0.450	0.081	0.000	0.218	0.086	0.011
Initial size	-0.030	0.090	0.739	-0.041	0.078	0.599	-0.022	0.101	0.828