Ann. Inst. Statist. Math.
Vol. 56, No. 1, 87–100 (2004)
©2004 The Institute of Statistical Mathematics

SMOOTHING ESTIMATION OF RATE FUNCTION FOR RECURRENT EVENT DATA WITH INFORMATIVE CENSORING

CHIN-TSANG CHIANG[1] AND MEI-CHENG WANG[2]

[1] *Department of Mathematics, National Taiwan University, Taipei 106, Taiwan, R.O.C.*
[2] *Department of Biostatistics, School of Public Health, The Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205-2179, U.S.A.*

(Received August 29, 2001; revised February 18, 2003)

Abstract. This paper proposes kernel estimation of the occurrence rate function for recurrent event data with informative censoring. An informative censoring model is considered with assumptions made on the joint distribution of the recurrent event process and the censoring time without modeling the censoring distribution. Under the validity of the informative censoring model, we also show that an estimator based on the assumption of independent censoring becomes inappropriate and is generally asymptotically biased. To investigate the asymptotic properties of the proposed estimator, the explicit form of its asymptotic mean squared risk and the asymptotic normality are derived. Meanwhile, the empirical consistent smoothing estimator for the variance function of the estimator is suggested. The performance of the estimators are also studied through Monte Carlo simulations. An epidemiological example of intravenous drug user data is used to show the influence of informative censoring in the estimation of the occurrence rate functions for inpatient cares over time.

Key words and phrases: Occurrence rate function, independent censoring, informative censoring, kernel estimator, longitudinal study, Poisson process.

1. Introduction

In biomedical and epidemiological longitudinal studies, recurrent event data are frequently collected from a group of subjects experiencing recurrent events. As commonly caused by loss to follow-up in a longitudinal study, the observation of each subject's recurrent events could be terminated before the end of study. Let $N(t)$ be the number of recurrent events over the time interval $[0, t]$ and Y be the censoring time (i.e., the time to the end of follow-up). Assume that the observation of the recurrent event process $N(t)$ is terminated at time Y. Generally, the recurrent event data are collected from n independent subjects. For the i-th subject, $N_i(t)$, Y_i and $\{t_{ij}\}_{j=1}^{m_i}$ respectively denote the recurrent event process, the censoring time, and the ordered event times observed in the time interval $[0, Y_i]$. Here, m_i is the number of the recurrent events occurring at or prior to Y_i.

In applications, because censoring could be caused by informative drop-out or death, it is sometimes unrealistic to assume the recurrent event process to be independent of the censoring time. To handle informative censoring or missing data in regression models, Robins *et al.* (1995) and Scharfstein *et al.* (1999) proposed semi-parametric methods to model the censoring distribution and adjust the estimation procedures via the censoring or missingness distribution function. In some applications, nonetheless, modeling the censoring mechanism could be intractable or undesirable because the censoring distribution is treated as a nuisance parameter in the model. In this paper, instead of modeling the censoring distribution, we consider a multiplicative intensity model which possesses appropriate interpretations and avoids modeling on the censoring distribution. Explicitly, the considered informative censoring model consists of the following assumptions.

(A1) Suppose there exists a nonnegative-valued latent variable Z_i so that, conditioning on z_i, $N_i(t)$ is a non-stationary Poisson process with the intensity function $\lambda_i(t) = z_i\phi_0(t)$, where $\phi_0(t)$ is a deterministic intensity function. The occurrence rate function of recurrent events in the population is $\lambda(t) = E[\lambda_i(t)] = \mu_z\phi_0(t)$ with $\mu_z = E[Z_i]$.

(A2) Conditioning on z_i, $N_i(\cdot)$ is independent of Y_i.

When the censoring time Y_i is independent of the recurrent event process $N_i(\cdot)$ unconditional on z_i, this informative censoring model reduces to an independent censoring model. The related works for the estimation of the cumulative rate function, which is formulated in the framework of counting process, can be tracked back to the publications of Nelson (1988) and Andersen *et al.* (1993). For the cumulative occurrence rate function without the Poisson assumption of (A1), the related works can be backed to the studies of Pepe and Cai (1993), Lawless and Nadeau (1995), and Lawless *et al.* (1997). Lin *et al.* (2000) and Sun and Wei (2000) also provided the non-parametric and semi-parametric estimators for the cumulative occurrence rate and the regression parameters with a rigorous justification through the modern empirical process theory.

In the above informative censoring model, the recurrent event process $N_i(\cdot)$ is assumed to be independent of the censoring time Y_i conditional on z_i, and is therefore allowed to be correlated with the censoring time Y_i through Z_i. Thus, the dependence between the recurrent events and the censoring time is modeled via the latent variable Z_i. It can be viewed as a non-parametric version of the random-effect model considered by Lancaster and Intrator (1998). In comparison with the Lancaster and Intrator model, note that the hazard function of Y_i and the distribution of Z_i are both left unspecified in our model. For related research which takes into account the covariate information, please refer to Wang *et al.* (2001).

The succeeding sections of this paper are organized as follows: In Section 2, kernel estimation method is proposed to estimate the occurrence rate function for the recurrent events with informative censoring. Here, a natural kernel estimator based on the assumption of independent censoring is also considered. The asymptotic properties of our estimator are studied in Section 3. Moreover, the empirical consistent smoothing estimator for the variance function of the estimator is suggested. It is shown in these two sections that the independent censoring estimator becomes inappropriate and is generally asymptotically biased when censoring is informative. Monte Carlo simulations are conducted in Section 4 to examine the performance of our estimators. In Section 5, the estimation procedure is applied to the intravenous drug user data. A brief discussion is provided in Section 6 and the proofs of the main results are placed in the Appendix.

2. Estimation

Consider the situation when the censoring time may be correlated with the recurrent event process. For example, patients of higher hospitalization frequency may be sicker and therefore drop out of the study earlier. As will be discussed later, an estimator based on the assumption of independent censoring becomes inappropriate and is generally asymptotically biased under the validity of the informative censoring model. This is caused because of the biased sampling of $\{Z_i\}$ in the risk sets. In this section, a method is proposed to avoid the influence of the latent variable Z_i in the kernel estimation.

Before introducing our estimation method, let $\Lambda_i(t)$, $\Phi_0(t)$ and $\Lambda(t)$ be the cumulative functions of $\lambda_i(t)$, $\phi_0(t)$ and $\lambda(t)$, respectively. Define the density function $f(t) = \phi_0(t)/\Phi_0(T_0)$, $0 \leq t \leq T_0$, as the normalized intensity function of $\phi_0(t)$ and let $F(t)$ be the distribution function of $f(t)$. In applications, the constant T_0 is usually selected as the maximum value of the observed censoring times $\{Y_i\}$, or as the maximum value of the event times $\{t_{ij}\}$. Under Assumption (A1), because

$$\frac{\lambda_i(t)}{\Lambda_i(T_0)} = \frac{z_i\phi_0(t)}{z_i\Phi_0(T_0)} = f(t),$$

the density function $f(t)$ can be viewed as the shape parameter for the intensity function $\lambda_i(t)$. Further, conditioning on (m_i, y_i, z_i), the event times $(t_{i1}, \ldots, t_{im_i})$ are the order statistics of a set of independent and identically distributed random variables with the truncated density $f(t)/F(y_i)$, $t \in [0, y_i]$. Thus, conditioning on $\{(m_i, y_i, z_i)\}$, the likelihood function is

$$L_c = \prod_{i=1}^{n}\left\{m_i!\prod_{j=1}^{m_i}\frac{f(t_{ij})}{F(y_i)}\right\} \propto \prod_{i=1}^{n}\prod_{j=1}^{m_i}\frac{f(t_{ij})}{F(y_i)}. \tag{2.1}$$

In the non-parametric models where ϕ_0 is an unspecified intensity function, the likelihood function L_c in (2.1) is essentially a non-parametric likelihood for right-truncated data. For the estimation of the distribution function $F(t)$, we can use the non-parametric maximum likelihood estimator $\widehat{F}(t)$, proposed by Wang *et al.* (2001), of the form

$$\widehat{F}(t) = \prod_{\{s_{(l)}>t\}}\left(1 - \frac{d_{(l)}}{N_{(l)}}\right),$$

where $\{s_{(l)}\}$ are the order statistics of the event times $\{t_{ij}\}$, $d_{(l)}$ is the number of events occurring at $s_{(l)}$, and $N_{(l)}$ is the total number of events with event time and censoring time satisfying $t_{ij} \leq s_{(l)} \leq y_i$.

The first step of our estimation procedure uses $(m_i^{-1}\sum_{j=1}^{m_i} K_{Y_i}(\frac{t-t_{ij}}{h}))$ to estimate $f(t)/F(Y_i)$, where $K_{Y_i}(\cdot)$ is a boundary kernel density of Gasser and Müller (1978) with adjustment for the censoring time Y_i, and h is a positive-valued bandwidth. Instead of estimating the subject intensity function $\lambda_i(t)$, we do this mainly to avoid the influence of latent variables $\{Z_i\}$. Due to the truncation of the recurrent event process $N_i(t)$ at Y_i, we multiply $(m_i^{-1}\sum_{j=1}^{m_i} K_{Y_i}(\frac{t-t_{ij}}{h}))$ by $\Lambda(Y_i) = F(Y_i)\Lambda(T_0)$, since $(f(t)/F(Y_i))\Lambda(Y_i) = \mu_z\phi_0(t) = \lambda(t)$. Because $F(t)$ and $\Lambda(T_0)$ are unknown, the third step is to substitute them with appropriate estimators. Conditional on (y_i, z_i), the number of the observed events, m_i, has the Poisson distribution with the expected value $z_i\Phi_0(y_i)$, we derive

$$\begin{aligned} E[m_i/F(Y_i)] &= E[E[m_i/F(Y_i) \mid (Y_i, Z_i)]] = E[Z_i\Phi_0(Y_i)/F(Y_i)] \\ &= E[Z_i\Phi_0(T_0)] = \Lambda(T_0). \end{aligned}$$

By substituting $\widehat{F}(t)$ for $F(t)$, an estimator for $\Lambda(t)$ can be constructed as

$$\widehat{\Lambda}(t) = \widehat{F}(t)\left(n^{-1}\sum_{l=1}^{n} m_l/\widehat{F}(Y_l)\right).$$

After the above adjustments, a kernel estimator is proposed as

$$\widetilde{\lambda}_h(t) = \sum_{i=1}^{n} \frac{\delta_i^*(t)}{\delta_\cdot^*(t)}\left(\frac{\widehat{\Lambda}(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right), \quad t \in [0, T_0], \tag{2.2}$$

where $\delta_i^*(t) = I(Y_i \geq t, m_i \geq 1)$ and $\delta_\cdot^*(t) = \sum_{i=1}^{n} \delta_i^*(t)$. In (2.2), with subject-specific estimators of intensity functions being appropriately formulated, the kernel estimator is essentially the average of subject-specific estimators defined with respect to each risk set at t.

When Y_i is independent of $N_i(\cdot)$ unconditional on z_i, the non-parametric kernel estimation method of Bartoszyński *et al.* (1981) can be extended to the following kernel estimaor

$$\widehat{\lambda}_h(t) = \sum_{i=1}^{n} \frac{\delta_i(t)}{\delta_\cdot(t)}\left(\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right), \quad t \in [0, T_0], \tag{2.3}$$

where $\delta_i(t) = I(Y_i \geq t)$ is an indicator function. The term $(\sum_{j=1}^{m_i} K_{Y_i}(\frac{t-t_{ij}}{h}))$ in (2.3) is a natural kernel estimator of the subject-specific intensity function $\lambda_i(t)$ for t in the interval $[0, Y_i]$. In the estimation, $\widehat{\lambda}_h(t)$ uses the information of subjects who are still at risk at t, i.e., $Y_i \geq t$. Because the censoring is independent of the recurrent event process, the risk set at each t forms a random sample from the population and the smoothing technique uses the risk set as the base for kernel estimation. However, in the next section, we will show that this independent censoring estimator is asymptotically biased under the assumptions of informative censoring. This is mainly caused by the biased sampling of $\{Z_i\}$ in the risk sets.

3. Asymptotic properties

In this section, the asymptotic risks of the kernel estimator $\widetilde{\lambda}_h(t)$ are established. The properties developed here are based on the mean squared error and the asymptotic normality of the estimator. Let $MSE(\widetilde{\lambda}_h(t))$ and $MISE(\widetilde{\lambda}_h) = \int MSE(\widetilde{\lambda}_h(t))\pi(t)dt$, where $\pi(t)$ is a non-negative weight function, represent the mean squared error and the mean integrated squared error of $\widetilde{\lambda}_h(t)$. By the decomposition principle of the mean squared error, it is convenient to consider the bias and the variance of $\widetilde{\lambda}_h(t)$, which are denoted separately by $B(\widetilde{\lambda}_h(t))$ and $V(\widetilde{\lambda}_h(t))$. Throughout this section, Y_i's are assumed to be independent and identically distributed with the cumulative distribution function $F_Y(y)$ and the probability measure $P_Y(y)$. Moreover, $P_{YZ}(y,z)$, $P_{mY}(k,y)$ and $P_{mYZ}(k,y,z)$ represent the probability measures of (Y,Z), (m,Y) and (m,Y,Z).

Before the derivation of the main results, the following conditions are assumed:

(A3) $G(t) = \int zI(y \geq t)dP_{YZ}(y,z)$ is a continuous function for $t \in [0, T_0]$.

(A4) The kernel density $K_s(\frac{t-u}{h}) = \frac{1}{h}\alpha(s, \frac{t-u}{h})K(\frac{t-u}{h})$ is continuous, bounded and satisfies

$$\beta_0(t,h,s) = 1, \quad \beta_1(t,h,s) = 0, \quad \beta_2(t,h,s) < \infty, \quad \text{and} \quad \gamma_4(t,h,s) < \infty,$$

where $\beta_j(t,h,s) = \int_{(t-s)/h}^{t/h} u^j \alpha(s,u)K(u)du$ and $\gamma_l(t,h,s) = \int_{(t-s)/h}^{t/h} |\alpha(s,u)K(u)|^l du$.

(A5) $\lambda(t)$ is twice differentiable and bounded.

(A6) $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

As one can see, the property of non-parametric estimator $\widehat{\Lambda}(t)$ will be used in the subsequent proofs for the asymptotic risks of $\widetilde{\lambda}_h(t)$. Define $Q(t) = \int_0^t G(u)\lambda(u)du$ and $R(t) = G(t)\Lambda(t)$. Wang *et al.* (2001) derived the i.i.d. representation of $\widehat{\Lambda}(t)$ as

$$\widehat{\Lambda}(t) = \Lambda(t)\left(1 + \frac{1}{n}\sum_{i=1}^{n} d_i(t) + o_p(n^{-1/2})\right), \tag{3.1}$$

where

$$d_i(t) = -\int \frac{kb_i(y)dP_{mY}(k,y)}{\Lambda(y)} + \frac{m_i}{\Lambda(Y_i)} - 1 + b_i(t)$$

with

$$b_i(t) = \sum_{j=1}^{m_i}\left(\int_t^{T_0} \frac{I(t_{ij} \le u \le Y_i)dQ(u)}{R^2(u)} - \frac{I(t < t_{ij} \le T_0)}{R(t_{ij})}\right).$$

Since $d_i(t)$ has zero expectation and the finite second moment $E[d_1^2(t)]$, they showed the asymptotic normality of $\widehat{\Lambda}(t)$ below.

LEMMA 3.1. *Suppose that assumptions* (A1)–(A3) *and* $F_Y(T_0) < 1$ *are satisfied. When* $n \to \infty$ *and* $t \in (0, T_0]$, $\sqrt{n}(\hat{\Lambda}(t) - \Lambda(t)) \xrightarrow{d} N(0, E[d_1^2(t)])$.

PROOF. See Wang *et al.* (2001). □

From (3.1), the estimator $\tilde{\lambda}_h(t)$ in (2.2) can be re-expressed as

$$\tilde{\lambda}_h(t) = \sum_{i=1}^{n} \xi_i(t)\left(1 + \frac{1}{n}\sum_{l=1}^{n} d_l(Y_i) + o_p(n^{-1/2})\right), \tag{3.2}$$

where $\xi_i(t) = \frac{\delta_i^*(t)}{\delta_\cdot^*(t)}\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right)$. Before deriving the mean squared error of $\tilde{\lambda}_h(t)$, we first state the following technical lemma. Under the regularity conditions, the moments of $\xi_i(t)$ can be obtained.

LEMMA 3.2. *Suppose that assumptions* (A1)–(A6) *and* $\mu_{\delta^*}(T_0) > 0$ *are satisfied. When n is sufficiently large and* $t \in (0, T_0]$,

$$E[\xi_i(t)] = \lambda(t)n^{-1} + b_\lambda(t)n^{-1}h^2 + o(n^{-1}h^2), \tag{3.3}$$

$$E[\xi_i^2(t)] = \sigma_\lambda^2(t)n^{-2}h^{-1} + o(n^{-2}h^{-1}) \tag{3.4}$$

and

$$E[|\xi_i(t)|^\nu] \le \left(\frac{\lambda(t)}{\mu_{\delta^*}^\nu(t)}\int_{\{k\ge 1, y \ge t\}} \frac{\gamma_\nu(t,h,y)\Lambda^{\nu-1}(y)}{k^{\nu-1}}dP_{mYZ}(k,y,z)\right) \times n^{-\nu}h^{-\nu+1}, \tag{3.5}$$

where $\nu = 3, 4$,

$$b_\lambda(t) = \left(\frac{\lambda^{(2)}(t)(\int_{\{k\geq 1, y\geq t\}} \beta_2(t,h,y)dP_{mYZ}(k,y,z))}{2\mu_{\delta^*}(t)} \right)$$

and

$$\sigma_\lambda^2(t) = \left(\frac{\lambda(t)(\int_{\{k\geq 1, y\geq t\}} (\gamma_2(t,h,y)\Lambda(y)/k)dP_{mYZ}(k,y,z))}{\mu_{\delta^*}^2(t)} \right)$$

with $\mu_{\delta^*}(t) = \int_{\{k\geq 1, y\geq t\}} dP_{mY}(k,y)$.

PROOF. See Appendix. □

From Lemma 3.2 and the relation between $\tilde{\lambda}_h(t)$ and $\sum_{i=1}^n \xi_i(t)$ in (3.2), we have the bias and the variance of $\tilde{\lambda}_h(t)$.

THEOREM 3.1. *Suppose that assumptions* (A1)–(A6) *and* $\mu_{\delta^*}(T_0) > 0$ *are satisfied. When* n *is sufficiently large and* $t \in (0, T_0]$, *the bias and the variance of* $\tilde{\lambda}_h(t)$ *are*

$$B(\widetilde{\lambda}_h(t)) = b_\lambda(t)h^2 + o(h^2) + O(n^{-1}h^{-1/2}) \tag{3.6}$$

and

$$V(\widetilde{\lambda}_h(t)) = \sigma_\lambda^2(t)(nh)^{-1} + o((nh)^{-1}). \tag{3.7}$$

PROOF. See Appendix. □

Under the validity of the informative censoring model, $\widetilde{\lambda}_h(t)$ is shown to be an asymptotically unbiased estimator of $\lambda(t)$. However, the estimator $\widehat{\lambda}_h(t)$ in (2.3) is generally asymptotically biased. When n is sufficiently large, the bias of $\widehat{\lambda}_h(t)$ is derived as follows:

$$B(\widehat{\lambda}_h(t)) = \lambda(t)\left(1 - \frac{G(t)}{\mu_z(1 - F_Y(t))}\right)(1 + o(1)). \tag{3.8}$$

The proof for $B(\widehat{\lambda}_h(t))$ is along with the same lines as the proof for $B(\widetilde{\lambda}_h(t))$. The asymptotic representations for $MSE(\widetilde{\lambda}_h(t))$ and $MISE(\widetilde{\lambda}_h)$ can be obtained from (3.6) and (3.7). It follows from assumption (A6) that both $B(\widetilde{\lambda}_h(t))$ and $V(\widetilde{\lambda}_h(t))$ converge to zero. This convergence rate now depends on whether and how the sample size converges to infinity and the bandwidth h converges to zero. As for the convergence rates of $MSE(\widetilde{\lambda}_h(t))$ and $MISE(\widetilde{\lambda}_h)$, the best convergence rates are equal to $n^{-4/5}$, which is attained by taking $h = O(n^{-1/5})$. When the further assumption, (A6′) $h = n^{-1/5}h_0$ for some positive bounded constant h_0, is made, we can derive

$$\sqrt{nh}\left(\frac{1}{n}\sum_{i=1}^n \xi_i(t) - \tilde{\lambda}_h(t)\right) \xrightarrow{p} 0, \quad \text{as} \quad n \to \infty. \tag{3.9}$$

From (3.9) and by using the Berry-Esséen theorem to the quantity $\sum_{i=1}^n \xi_i(t)$, the asymptotic normality of $\tilde{\lambda}_h(t)$ is then obtained in the following theorem.

THEOREM 3.2. *Suppose that assumptions* (A1)–(A5), (A6′) *and* $\mu_{\delta^*}(T_0) > 0$ *are satisfied. When* n *converges to infinity,*

$$\sup_u \left| P\left(\frac{\sqrt{nh}(\tilde{\lambda}_h(t) - \lambda(t)) - b_\lambda(t)}{\sigma_\lambda(t)} \le u \right) - \Phi(u) \right| \to 0, \tag{3.10}$$

where $\Phi(\cdot)$ *denotes the cumulative distribution function of the standard normal distribution.*

In the above derivation, it is found that $\sqrt{\frac{h}{n}} \sum_{i=1}^n \xi_i(t)$ and $\sqrt{nh}\tilde{\lambda}_h(t)$ have the same asymptotic variance. Thus, by using the i.i.d. property of $\xi_i(t)$'s and substituting the non-parametric estimator $\widehat{\Lambda}(t)$ for $\Lambda(t)$, we propose the empirical consistent smoothing estimator of $V(\tilde{\lambda}_h(t))$ by

$$\widehat{V}(\tilde{\lambda}_h(t)) = \frac{1}{\delta_\cdot^{*2}(t)} \sum_{i=1}^n \delta_i^*(t) \left(\frac{\widehat{\Lambda}(Y_i)}{m_i} \sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t - t_{ij}}{h}\right) - \tilde{\lambda}_h(t) \right)^2 . \tag{3.11}$$

With similar arguments as the derivation of Lemmas 3.1, 3.2 and (3.9), the quantity $\sqrt{nh}\widehat{V}(\tilde{\lambda}_h(t))$ can be shown to converge to $\sigma_\lambda^2(t)$ in probability.

4. Monte Carlo simulation

To examine the finite sample properties of $\tilde{\lambda}_h(t)$ and the empirical smoothing estimator for the variance function of $\tilde{\lambda}_h(t)$, Monte Carlo simulations are used. The simulated data are similar in nature to those recurrent event data collected in empirical biomedical and epidemiological cohort studies. Two types of recurrent event data are generated below.

Let the latent variables Z_i be independent and identically distributed with the uniform distribution $U(0.5, 4)$. The first data set is generated from 400 independent non-stationary Poisson processes $\{N_i(t)\}$ with the corresponding subject-specific intensity functions $\lambda_i(t) = z_i \phi_0(t)$, where

$$\phi_0(t) = 3 + \frac{(t-6)^3}{72}, \qquad t \in [0, 10].$$

Since the expectation of Z is equal to 2.25, it implies that the occurrence rate function $\lambda(t) = 2.25\phi_0(t)$ for $t \in [0, 10]$. Conditional on z_i, the censoring time Y_i is designed to be distributed as a truncated distribution of the exponential distribution $\exp(z_i/10)$, where the truncated distribution ranges from 1 to 10 and has the density

$$f_{Y|z_i}(y) = \frac{0.1 z_i \exp(-0.1 z_i y)}{(\exp(-0.1 z_i) - \exp(-z_i))}, \qquad y \in [1, 10].$$

This simulated data set satisfies the assumptions of informative censoring. When z_i in $\exp(z_i/10)$ is replaced by a constant, the informative censoring condition in (A2) then reduces to the independent censoring condition. The second data set is generated so that it is similar to the first except that z_i in $\exp(z_i/10)$ is substituted by 2.25.

Computed by (2.1) and (2.3), the kernel estimators $\tilde{\lambda}_h(t)$ and $\widehat{\lambda}_h(t)$ are applied to the simulated data. Moreover, the empirical consistent smoothing estimators of $V(\tilde{\lambda}_h(t))$ in

(3.11) and $V(\tilde{\lambda}_h(t))$ are provided. Here, the considered smoothing estimator of $V(\tilde{\lambda}_h(t))$ is suggested by

$$\widehat{V}(\tilde{\lambda}_h(t)) = \frac{1}{\delta_{\cdot}^2(t)} \sum_{i=1}^{n} \delta_i(t) \left(\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t - t_{ij}}{h}\right) - \widehat{\lambda}_h(t) \right)^2 . \tag{4.1}$$

For these estimators, an appropriate bandwidth is used by examining the plots of the estimated curves, and the Gaussian kernel is selected for $K(\cdot)$. Based on other kernels, such as the Epanechnikov kernel and the uniform kernel, estimators gave similar results, hence, are omitted. Also, alternative selections of bandwidth are also possible. Figures 1a and 1b show the true occurrence rate function $\lambda(t)$, the 1000 simulation averages of the estimates $\tilde{\lambda}_h(t)$, $\widehat{\lambda}_h(t)$ and their corresponding estimated ± 1.96 standard bars, and the ± 1.96 standard errors of the 1000 occurrence rate estimates at the corresponding time points. As shown in Fig. 1a, there is no observable difference between these two estimates. However, in Fig. 1b, $\tilde{\lambda}_h(t)$ provides a much less biased estimate under assumptions of the informative censoring model. The similarity and difference in the estimated rate functions are clearly due to the absence or presence of informative censoring. Meanwhile, it can be found in these figures that the estimated standard errors are very close to the true standard errors of the estimators.

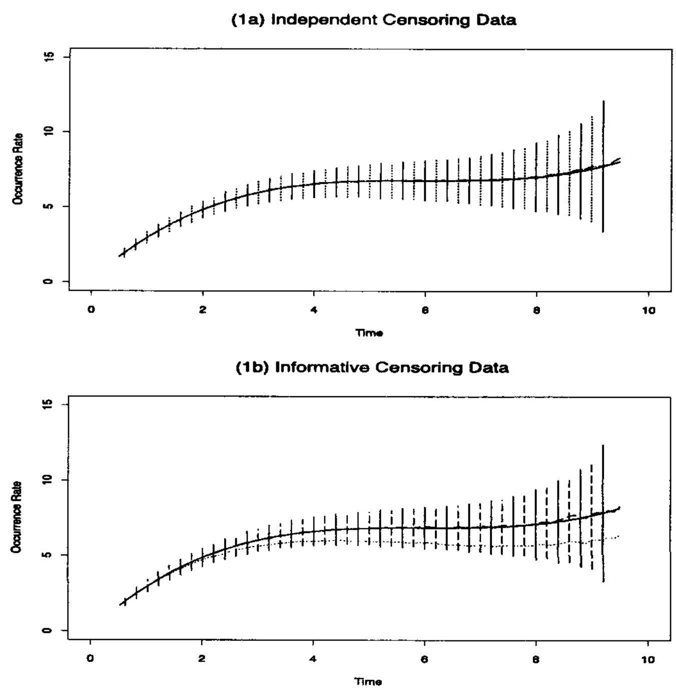

Fig. 1. The real occurrence rate function $\lambda(t)$ (solid curve) and the estimated occurrence rate functions $\tilde{\lambda}_h(t)$ (dashed curve) and $\widehat{\lambda}_h(t)$ (dotted curve). (1a) The ± 1.96 standard errors (solid line) and the estimated ± 1.96 standard errors of $\widehat{\lambda}_h(t)$ (dotted line) at the corresponding time points for recurrent event data with independent censoring. (1b) The ± 1.96 standard errors (solid line) and the estimated ± 1.96 standard errors of $\tilde{\lambda}_h(t)$ (dashed line) at the corresponding time points for recurrent event data with informative censoring.

5. A data example

The data from the AIDS Link to Intravenous Experiences cohort study (Vlahov *et al.* (1991)) provides information on inpatient admissions and Human Imunodeficiency Virus (HIV) status of intravenous drug users. In total, there are 297 HIV-positive and 450 HIV-negative intravenous drug users involved in this study. The study was initiated in 1988 and started to systematically collect health service data in July, 1993. The repeated hospitalizations for each drug user here were observed between August 1, 1993 and December 31, 1997. Let t_{ij}, $j = 1, \ldots, m_i$, be the time length from August, 1, 1993 to the date of the j-th inpatient admission, y_i the time length to the last visit for the i-th drug user, and T_0 the maximum time of y_i's. In our analysis, we consider only drug users who entered the study before July 16, 1993. A drug user is defined to be HIV-positive if he or she was infected by HIV-1 virus prior to July 16, 1993. A drug user is defined to be HIV-negative if the individual was not recorded as HIV-positive at anytime before December 31, 1997. Those drug users whose HIV-1 infection occurred in the study period are excluded from our data analysis.

Among HIV-positive drug users, the median of the number of recurrent events is 2 and the number ranges from 0 to 14. The mean of the censoring time is 3.257 years and the censoring time ranges from 0.047 to 4.394 years. For HIV-negative drug users, the

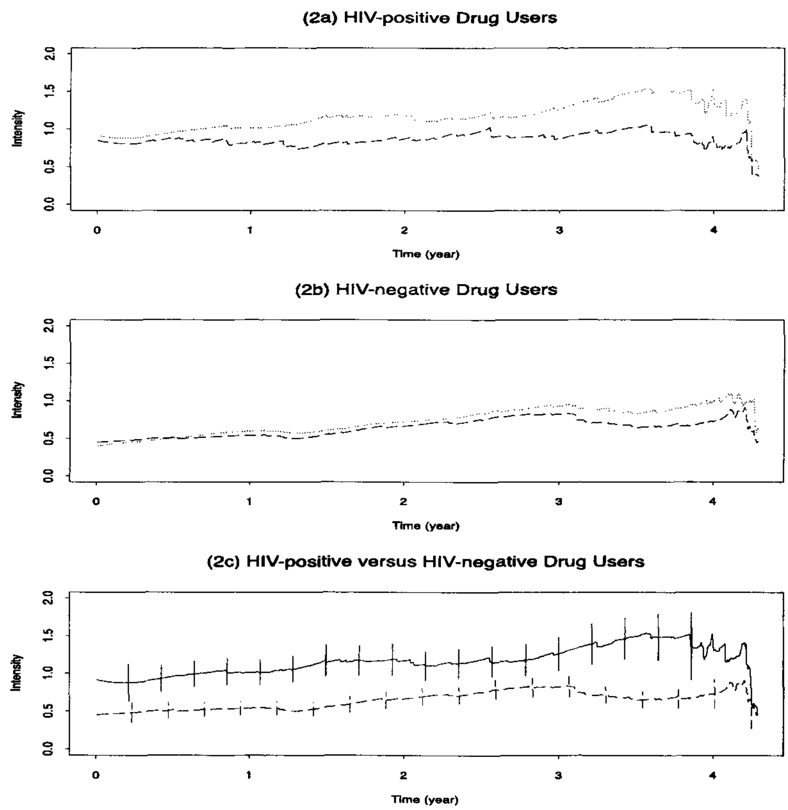

Fig. 2. The estimated occurrence rate functions $\widetilde{\lambda}_h(t)$ (solid curve) and $\widehat{\lambda}_h(t)$ (dashed curve) of HIV positive and HIV negative intravenous drug users.

median of the number of recurrent events is 1 and the number ranges from 0 to 19. The mean of the censoring time is 3.734 and the censoring time ranges from 0.275 to 4.394 years. The main objective of this study is to estimate the occurrence rate functions of hospitalizations for HIV-positive and HIV-negative drug users. The data serves as a good example for our methodology because the drug users are likely to drop out for reasons associated with the outcome measurement and, in AIDS research, the HIV-positive drug users are known to have high mortality rate which is apparently associated with the inpatient care measurement.

As in the simulation study, the kernel estimators $\widetilde{\lambda}_h(t)$ and $\widehat{\lambda}_h(t)$ are computed using an appropriate bandwidth and the Epanechnikov kernel density for $K(\cdot)$. Figures 2a and 2b show the estimated occurrence rate functions, which are computed by (2.2) and (2.3), for HIV-positive and HIV-negative drug users. Figure 2c further provides the corresponding ± 1.96 estimated standard error bars of the hospitalization rate estimators at the selected time points. We can see that the two different kernel estimates result in similar curves for HIV-negative drug users as shown in Fig. 2b. In contrast, in Fig. 2a, the two estimates result in very different hospitalization rate functions for HIV-positive drug users, suggesting the presence of significant informative censoring. Furthermore, the estimate $\widetilde{\lambda}_h(t)$ for HIV-positive drug users appears to be significantly higher than $\widehat{\lambda}_h(t)$ for HIV-negative drug users (Fig. 2c). The analysis essentially reveals the actual need of health service or insurance from HIV-positive drug users, an implication which cannot be derived from the descriptive statistics that are typically used in AIDS research.

6. Discussion

In this paper, we propose kernel estimation of the occurrence rate function for recurrent event data in an informative censoring model. The informative censoring model is constructed by assuming independence between the censoring time Y_i and the recurrent event process $N_i(\cdot)$ conditional on z_i. The multiplicative relationship between the latent variable Z_i and the baseline intensity function $\phi_0(t)$ in (A1) is the key assumption which makes the construction of $\widetilde{\lambda}_h(t)$ possible. As part of the requirement of the model, all the subject-specific occurrence rate functions of recurrent events are assumed to have the same baseline intensity function. Further research will involve work to consider alternative informative censoring models and also to develop methods to allow for time-dependent latent variable $Z_i(t)$.

Acknowledgements

The first author's research was partially supported by the National Science Council grant 91-2118-M-002-002 (Taiwan), and the second author's research was partially supported by the National Institute of Health grants R01 HD38209 and R01 MH56639 (U.S.A.). We would like to thank the referee for valuable comments. We are also grateful to Dave Vlahov and Strathdee at Johns Hopkins University for providing the anonymous ALIVE data. Provision of the data was supported by National Institute on Drug Abuse grants DA04334 and DA08009.

Appendix

PROOF OF LEMMA 3.2. By the definition of $\xi_i(t)$, it can be derived that

$$(A.1)\qquad E[\xi_i(t)] = E\left[\frac{\delta_i^*(t)}{\delta_\cdot^*(t)}\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right)\right]$$
$$= E[(\delta_\cdot^{*(i)}(t)+1)^{-1}]E\left[\delta_i^*(t)\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right)\right],$$

where $\delta_\cdot^{*(i)}(t) = \sum_{j\neq i}\delta_j^*(t)$. Since $\delta_\cdot^{*(i)}(t)$ is distributed as Binomial $(n-1, \mu_\delta^*(t))$, the expectation of $(\delta_\cdot^{*(i)}(t)+1)^{-1}$ is directly calculated as

$$(A.2)\qquad E[(\delta_\cdot^{*(i)}(t)+1)^{-1}] = (1-(1-\mu_\delta^*(t))^n)(n\mu_\delta^*(t))^{-1}.$$

Thus, (A.1) can be written as

$$(A.3)\qquad E[\xi_i(t)] = E\left[\delta_i^*(t)\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right)\right](n\mu_\delta^*(t))^{-1}(1+o(1)).$$

Let t_{ij}^* be unordered observations of t_{ij}. It implies from assumption (A1) that conditioning on (m_i, y_i, z_i), t_{ij}^* are independently identically distributed random variable with density $\lambda(t)/\Lambda(y_i)$ for $t \in [0, y_i]$. By assumptions (A1)–(A6) and the Taylor expansion, we can get

$$(A.4)\qquad E\left[\delta_i^*(t)\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right)\right]$$
$$= E\left\{E\left[\frac{\delta_i^*(t)\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right) \mid (m_i, Y_i, Z_i)\right]\right\}$$
$$= E\left\{E\left[\frac{\delta_i^*(t)\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}^*}{h}\right) \mid (m_i, Y_i, Z_i)\right]\right\}$$
$$= \int_{\{k\geq 1, y\geq t\}} \Lambda(y)\left(\int_0^y K_y\left(\frac{t-u}{h}\right)\frac{\lambda(u)}{\Lambda(y)}du\right)dP_{mYZ}(k,y,z)$$
$$= \mu_{\delta^*}(t)\lambda(t) + \left(\frac{\lambda^{(2)}(t)}{2}\int_{\{k\geq 1, y\geq t\}}\beta_2(t,h,y)dP_{mYZ}(k,y,z)\right)h^2 + o(h^2).$$

Substituting (A.4) into (A.3), the expectation of $\xi_i(t)$ in (3.3) is then obtained. From (A.2), it is straightforward to derive that

$$(A.5)\qquad E[\xi_i^2(t)] = E\left[\delta_i^*(t)\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right)^2\right]\frac{(1+o(1))}{(n^2\mu_{\delta^*}^2(t))}$$

with

$$E\left[\delta_i^*(t)\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i}K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right)^2\right] \tag{A.6}$$

$$= E\left\{E\left[\delta_i^*(t)\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i}K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right)^2 \mid (m_i, Y_i, Z_i)\right]\right\}$$

$$= E\left\{E\left[\delta_i^*(t)\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i}K_{Y_i}\left(\frac{t-t_{ij}^*}{h}\right)\right)^2 \mid (m_i, Y_i, Z_i)\right]\right\}$$

$$= (I + II),$$

where

$$I = E\left\{E\left[\frac{\delta_i^*(t)\Lambda^2(Y_i)}{m_i^2}\sum_{j=1}^{m_i}K_{Y_i}^2\left(\frac{t-t_{ij}^*}{h}\right) \mid (m_i, Y_i, Z_i)\right]\right\} \tag{A.7}$$

$$= \int_{\{k\geq 1, y\geq t\}} \frac{\Lambda(y)}{k}\left(\int_0^y K_y^2\left(\frac{t-u}{h}\right)\lambda(u)du\right) dP_{mYZ}(k,y,z)$$

$$= \left(\lambda(t)\int_{\{y\geq t, k\geq 1\}}(\gamma_2(t,h,y)\Lambda(y)/k)dP_{mYZ}(k,y,z)\right)h^{-1} + o(h^{-1})$$

and

$$II = E\left\{E\left[\frac{\delta_i^*(t)\Lambda^2(Y_i)}{m_i^2} \times \left(\sum_{j_1\neq j_2} K_{Y_i}\left(\frac{t-t_{ij_1}^*}{h}\right)K_{Y_i}\left(\frac{t-t_{ij_2}^*}{h}\right)\right) \mid (m_i, Y_i, Z_i)\right]\right\} \tag{A.8}$$

$$= \int_{\{k\geq 1, y\geq t\}} (1-k^{-1})\left(\int_0^y K_y\left(\frac{t-u}{h}\right)\lambda(u)du\right)^2 dP_{mYZ}(k,y,z)$$

$$= \left(\int_{\{k\geq 1, y\geq t\}}(1-k^{-1})dP_{mYZ}(k,y,z)\right)\lambda^2(t) + o(1).$$

Therefore, (3.4) is obtained from (A.5), (A.7) and (A.8). Along the same lines as the derivation of $E[\xi_i^2(t)]$, the statement in (3.5) follows.

PROOF OF THEOREM 3.1. From (3.3), (3.4) and the property $E[d_l(t)] = 0$, we get

$$E\left[\sum_{i=1}^n \xi_i(t)\right] = \lambda(t) + b_\lambda(t)h^2 + o(h^2) \tag{A.9}$$

and

$$
\begin{aligned}
\text{(A.10)} \qquad & E\left[\sum_{i=1}^{n} \xi_i(t)\left(\frac{1}{n}\sum_{l=1}^{n} d_l(Y_i)\right)\right] \\
&= E\left[\sum_{i=1}^{n} \xi_i(t)\left(\frac{1}{n}\sum_{l\neq i}^{n} d_l(Y_i)\right)\right] + \frac{1}{n}E\left[\sum_{i=1}^{n} \xi_i(t) d_i(Y_i)\right] \\
&= E\left[\sum_{l\neq i}^{n} d_l(Y_i)\delta_i^*(t)\left(\frac{\Lambda(Y_i)}{m_i}\sum_{j=1}^{m_i} K_{Y_i}\left(\frac{t-t_{ij}}{h}\right)\right)\right]\frac{(1+o(1))}{\mu_{\delta^*}(t)} \\
&\quad + O(n^{-1}h^{-1/2}) \\
&= n\left(\int E[d_l(y)]\left(\int_0^y K_y\left(\frac{t-u}{h}\right)\lambda(u)du\right) dP_{mYZ}(k,y,z)\right) \\
&\quad \times \frac{(1+o(1))}{\mu_{\delta^*}(t)} + O(n^{-1}h^{-1/2}) \\
&= O(n^{-1}h^{-1/2}).
\end{aligned}
$$

From (A.8) and (A.9), the bias of $\tilde{\lambda}_h(t)$ in (3.6) is derived. The proof of (3.7) is developed by considering $E[\tilde{\lambda}_h^2(t)]$. Straightforward decomposition shows that

$$
\text{(A.11)} \qquad E[\tilde{\lambda}_h^2(t)] = III + IV,
$$

with

$$
III = E\left[\sum_{i=1}^{n} \xi_i^2(t)\left(1 + \frac{2}{n}\sum_{l=1}^{n} d_l(Y_i) + o_p(n^{-1/2})\right)\right]
$$

and

$$
IV = (n^2-n)E\left[\prod_{i=1}^{2}\left(\xi_i(t)\left(1+\frac{1}{n}\sum_{l=1}^{n} d_l(Y_i) + o_p(n^{-1/2})\right)\right)\right].
$$

From (3.4)–(3.5), it follows that

$$
\text{(A.12)} \qquad E\left[\sum_{i=1}^{n} \xi_i^2(t)\left(\frac{1}{n}\sum_{l=1}^{n} d_l(Y_i)\right)\right] = O(n^{-2}h^{-3/2})
$$

and, hence,

$$
\text{(A.13)} \qquad III = \sigma_\lambda^2(t)(nh)^{-1} + o((nh)^{-1}).
$$

Similarly, we can show that

$$
\text{(A.14)} \qquad IV = (E[\tilde{\lambda}_h(t)])^2(1+o(1)).
$$

Substituting (A.13) and (A.14) into (A.11), the variance in (3.7) is than derived.

References

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer, New York.

Bartoszyński, R., Brown, B. W., McBride, C. M. and Thompson, J. R. (1981). Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary poisson process, *Ann. Statist.*, **9**, 1050–1060.

Gasser, Th. and Müller, H.-G. (1978). Kernel estimation of regression functions, *Smoothing Techniques for Curve Estimation* (eds. Th. Gasser and M. Rosenblatt), Lecture Notes in Mathematics, No. 757, 23–68, Springer, Berlin.

Lancaster, T. and Intrator, O. (1998). Panel data with survival: Hospitalization of HIV-positive patients, *J. Amer. Statist. Assoc.*, **93**, 46–53.

Lawless, J. F. and Nadeau, C. (1995). Some simple robust method for the analysis of recurrent events, *Technometrics*, **37**, 158–168.

Lawless, J. F., Nadeau, C. and Cook, R. J. (1997). Analysis of mean and rate functions for recurrent events, *Proceedings of the First Seattle Symposium in Biostatistics*: *Survival Analysis* (eds. D. Y. Lin and T. R. Fleming), 37–49, Springer, New York.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events, *J. Roy. Statist. Soc. Ser. B*, **62**, 711–730.

Nelson, W. B. (1988). Graphical analysis of system repair data, *Journal of Quality Technology*, **20**, 24–35.

Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates, *J. Amer. Statist. Assoc.*, **88**, 811–820.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *J. Amer. Statist. Assoc.*, **90**, 106–121.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models, *J. Amer. Statist. Assoc.*, **94**, 1096–1120.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times, *J. Roy. Statist. Soc. Ser. B*, **62**, 293–302.

Vlahov, D., Anthony, J. C., Muñov, A., Margolick, J., Nelson, K. E., Celentano, D. D., Solomon, L. and Polk, B. F. (1991). The ALIVE study: A longitudinal study of HIV-1 infection in intravenous drug users: Description of methods, *The Journal of Drug Issues*, **21**, 759–776.

Wang, M. C., Qin, J. and Chiang, C. T. (2001). Analyzing recurrent event data with informative censoring, *J. Amer. Statist. Assoc.*, **96**, 1057–1065.