Nonparametric Estimation of Cumulative cause Specific Reversed Hazard Rates under Masked Causes of Failure

Sankaran PG[*] and Anjana S

Department of Statistics, Cochin University of Science and Technology, India

[*]Corresponding author: Sankaran PG, Department of Statistics, Cochin University of Science and Technology, Cochin-22, India, E-mail: sankaranpg@yahoo.com

Abstract

In the analysis of competing risks data, it is common that the exact cause of failure for certain study subjects is missing. This problem of missing failure type may be due to inadequacy in the diagnostic mechanism or reluctance to report the exact cause of failure. In the present paper, we consider the nonparametric estimation of cumulative cause specific reversed hazard rates for left censored competing risks data under masked causes of failure. We first develop maximum likelihood estimators of cumulative cause specific reversed hazard rates. We then consider the least squares type estimators for cumulative cause specific reversed hazard rates, when the information about the conditional probability of exact failure type given a set of possible failure types is available. Simulation studies are conducted to assess the performance of the proposed estimators. We illustrate the applicability of the proposed methods using a data set.

Keywords: Cause specific reversed hazard rates; Masked causes of failure; Weighted least squares estimation; Nonparametric estimation

Introduction

In survival studies, the failure (death) of subjects may be attributed to one of several causes or types, known as competing risks. In such situations, the subject is exposed to two or more causes of failure, but its eventual death can be due to exactly one of these causes of failure. In this context, for each subject, we observe a random vector (T, J) where T denotes (possibly censored) survival time and J represents cause of death (exactly one of k possible causes). J takes on values in the set $\{1, 2, ..., k\}$. Modeling and analysis of such competing risks data under right censoring are usually carried out using cause specific hazard rates $\lambda_j(t)$ defined by,

$$\lambda_J(t) = \lim_{\Delta t \to 0} \left(\frac{P\left[T < t + \Delta t, J = j \mid T \geq t\right]}{\Delta t} \right) \quad j = 1, 2, \ldots k. \quad (1.1)$$

For various properties of $\lambda_j(t)$, [1-3].

There are situations in the analysis of competing risks data where the exact cause of failure for certain subjects is missing. For example, due to inadequacy in the diagnostic mechanism, the experimentalists quite often are uncertain about the true failure type or are reluctance to report any specific value of J for some objects. Dinse was among the first to discuss the uncertainty in the information on failure types while estimating survival due to different failure types [4]. In such contexts, information on failure type is either completely available or not available at all. This problem with two failure types was studied subsequently by Miyakawa (1984), Racine - Poon and Hoel (1984), Lo (1991), Mukerjee and Wang (1993), Goetghebuer and Ryan (1990, 1995), Dewanji (1992) and Lu and Tsiatis (2001) [5-12]. Flehinger, et al. (1998) have considered a general pattern of missing failure types for the purpose of estimating survival due to different types, with the strong assumption of proportional hazards due to different types [13]. Flehinger, et al. (1998) emphasized on the parametric modelling with more general case, where the competing risks are not assumed to have proportional hazard functions [14]. Later Dewanji and Sengupta (2003) developed nonparametric maximum likelihood estimator for $\lambda_j(t)$ with missing at random assumption and also proposed a nonparametric estimator for cumulative cause specific hazard rates using counting process approach [15]. Recently, Sen, et al. proposed a semiparametric Bayesian approach for analyzing competing risks survival data with masked cause of death [16]. Hyun, et al. developed a semiparametric proportional hazards model for the cause specific hazard function in the analysis of competing risks data with missing cause of failure [17-20].

In survival studies, it is likely that some study objects encounter the event of interest before the start of the study. For example, in alzheimer studies, an elder person is at risk of multiple events like dementia and death. However, some subjects are demented before the start of the study, which is an incidence of left censoring. Thus, such studies give rise to left censored competing risks data. The existing models and methodologies for the analysis of competing risks data become inadequate in the presence of left censored observations.

The concept referred as reversed hazard rate, defined

$$h(t) = \lim_{\Delta t \to 0} \left(\frac{P\left[t - \Delta t < T \leq t \,|\, T \leq t \right]}{\Delta t} \right), \qquad (1.2)$$

plays a pivotal role in modeling and analysis of left censored failure time data. The function $h(t)$, which was proposed as a dual to the hazard rate by Barlow (1963), used in many contexts. In parallel systems of independent and identically distributed components, the hazard rate of the system lifetime is not proportional to the hazard rate of the lifetime of each component, however the reversed hazard rate of the system lifetime is proportional to the reversed hazard rate of the lifetime of the each component [21]. For various properties and applications of (1.2), [3,22-28].

Sankaran and Anjana presented the analysis of competing risks data using cause specific reversed hazard rates under left censoring [29]. Recently, Sankaran and Anjana introduced a proportional cause specific reversed hazards model for modeling and analysis of left censored competing risks data in the presence of covariates [30]. The problem of missing failure types was studied in literature by various researchers under right censoring [11,12,18-20]. Very Recently, Dewanji, *et al.* considered the regression problem, in which the cause specific hazard rates may depend on some covariates, and consider estimation of the regression coefficients and the cause specific baseline hazards under the general missing pattern using some semi- parametric models [31]. In many occasions, this problem of missing failure type may also arise under left censoring. Motivated by this, in this paper, we present nonparametric inference procedures for left censored competing risks data when causes of failure are masked.

The present paper is organized as follows. We defined the cause specific reversed hazard rates and study their properties, then discussed the nonparametric estimation of the cumulative cause specific reversed hazard rates under masking. We first formulated the likelihood function and considered the maximum likelihood estimation procedure for the estimation of cause specific reversed hazard rates. We then give least squares type estimator for cumulative cause specific reversed hazard rates, then the simulation studies are carried out to investigate the performance of the estimators. Then we applied the proposed procedures to a data set. Finally the conclusion includes our present work.

Cause specific reversed hazard rates

Let (T, J) be a pair of random variables as described in introduction Section . Let $F(t)$ be the distribution function of T. The cause specific reversed hazard rate of T is defined as

$$h_j(t) = \lim_{\Delta t \to 0} \left(\frac{P\left[t - \Delta t \leq T \leq t, \ J = j \,|\, T \leq t \right]}{\Delta t} \right) \quad j = 1, 2, ..., k. \qquad (2.1)$$

The $h_j(t)$ specifies the instantaneous rate of failure of a subject at time t due to cause j given that it failed before time [29]. Denote $F_j(t) = P[T \leq t, J = j]$ as the cumulative incidence function of T. We can write (2.1) as

$$h_j(t) = \frac{f_{j(t)}}{F(t)} \quad j = 1, 2, ..., k. \qquad (2.2)$$

where $f_j(t) = \dfrac{dF_j(t)}{dt}$ is the cause specific density of T. We assume that the k failures are mutually exclusive and exhaustive so that a subject can have at most one realized failure time with an identifiable cause. Then marginal reversed hazard rate for T is given by,

$$h(t) = \sum_{j=1}^{k} h_j(t)$$

Now the distribution function for T can be expressed in terms of cause specific reversed hazard rates as,

$$F(t) = exp\left[-H(t) \right] = exp\left[-\sum_{j=1}^{k} H_j(t) \right]. \qquad (2.3)$$

where $H(t) = \int_t^\infty h(u)du$ is the cumulative reversed hazard rate for T and $H_j(t) = \int_t^\infty h_j(u)du$. Now from (2.2), we obtain

$$F_j(t) = \int_0^t h_j(u)F(u)du = -\int_0^t F(u)dH_j(u) \qquad (2.4)$$

The function $h_j(t)$ fully describes the distribution of (T, J) in multiple failure mode settings. For more inference on (2.1) [29].

Our objective is to develop nonparametric estimators for cumulative cause specific reversed hazard rates and cumulative incidence functions for left censored competing risks data, when some of the causes of failure are masked.

Nonparametric estimation

In this section we discussed the nonparametric estimation of cumulative cause specific reversed hazard rates and cumulative incidence functions.

Let X be a non-negative random variable with distribution function $F(t)$ which is left censored by the random variable C . Under left censoring one could observe (T, δ), where $T = max(X, C)$ and δ is the censoring indicator(1 for failure and 0 for censoring).

In addition, we observe the set $G \subseteq \{1, 2, \ldots, k\}$ representing the possible failure types when $\delta = 1$. If failure occurs, G gives the partial information about the failure type. This information is complete when G is a singleton set.

In (Maximum likelihood estimator), we develop EM algorithm for nonparametric maximum likelihood estimation. In (Weighted least squares estimator), we suggest another procedure for estimation of cumulative cause of specific reversed hazard rates using method of weighted least squares. This method facilitates the analysis when information about the probability that a particular cause is responsible for the failure from a given set of possible causes is available to us.

Maximum likelihood estimator

Denote the observed data for i^{th} individual by $(t_i, \delta_i, \delta_i g_i)$, where t_i, δ_i, g_i are the observed values of T, δ, G respectively for $i = 1, 2\ldots n$. Assume that the missing at random assumption for observing g_i (Little and Rubin (1987)) that is given failure time and failure type, probability of observing g is same for all the types contained in g (Dewanji and Sengupta (2003)) [15,32]. The likelihood function for the observed data can be written as,

$$L = \prod_{i=1}^n \prod_{r \in A_i} \left\{ \left(\sum_{j \in g_r} h_j(t_i) \right)^{\delta_r} F(t_i) \right\} \qquad (3.1)$$

where A_i is the set of individuals failed or censored at time t_i. From (3.1) we can see that the nonparametric maximum likelihood estimators for the cause specific reversed hazard rates have masses at m distinct observed failure times $s_1 < s_2 < \ldots < s_m$. Then, we can write $h_j(s_i)$ as the discrete cause specific reversed hazard rate of type j at time s_i. Thus, using the identity

$$F(t) = \prod_{i:t_i>t} \left[1 - \sum_{j=1}^k h_j(t_i) \right], \text{3.1 becomes } L = \prod_{i=1}^m \left[\left(\prod_{r \in D_i} \sum_{j \in g_r} h_j(s_i) \right) \left(1 - \sum_{j=1}^k h_j(s_i) \right)^{n_i - d_i} \right] \qquad (3.2)$$

where D_i is the set of individuals failed at time s_i, d_i is the number of individuals failed at time s_i and n_i be the number of individual failed up to time s_i. We use EM algorithm for finding the maximum likelihood estimator for $h_j(t)$. By assuming that the cause of failure of each individual is available, the complete data likelihood can be written as,

$$L_c = \prod_{i=1}^m \left[\left\{ \prod_{r \in D_i} \left(\prod_{j=1}^k h_j(s_i)^{x_{jr}} \right) \right\} \left(1 - \sum_{j=1}^k h_j(s_i) \right)^{n_i - d_i} \right] \qquad (3.3)$$

$$= \prod_{i=1}^m \left[\left(\prod_{j=1}^k h_j(s_i)^{d_{ji}} \right) \left(1 - \sum_{j=1}^k h_j(s_i) \right)^{n_i - d_i} \right] \qquad (3.4)$$

Where

$x_{jr} = 1$ if rth individual failed due to cause j
$= 0$ otherwise,

$d_{ji} = \sum_{r \in D_i} x_{jr}$, which is the number of individuals failed due to cause j at time s_i and $d_i = \sum_{j=1}^{k} d_{ji}$.

The E step of the algorithm takes the conditional expectation of $d_{ji}'s$ or $x_{jr}'s$, given the initial estimate of $h_j(t_i)'s$ and the incomplete observed data. Thus the conditional expectation of x_{jr}, denoted by $x_{jr}^{(0)}$ is given as $x_{jr}^{(0)} = \dfrac{h_j^{(0)}(t_i)}{\sum_{j' \in g_r} h_{j'}^{(0)}(t_i)}$ for $j \in g_r$ and 0 otherwise. Then the conditional expectation of d_{ji} is obtained as $d_{ji}^{(0)} = \sum_{r \in D_i} x_{jr}^{(0)}$. The M step maximizes the conditional expectation of $logLc$ with respect to $h_j^{(0)}(t_i)$ to get the better estimates $h_j^{(1)}(t_i) = \dfrac{d_{ji}^{(0)}}{n_i}$. The process is repeated until the estimate converges.

Following the steps of Dewanji and Sengupta (2003), we can show that the observed information matrix corresponding to (3.2) is positive definite and then the likelihood function (3.2) is concave and has a unique maximum [15]. Thus the EM algorithm mentioned above converges to this unique maximum (Dempster *et al.* (1977), Wu (1983)) [33,34].

The above mentioned EM algorithm gives the nonparametric estimators, $\hat{h}_j(t)$ for cause of specific reversed hazard rates. Thus the nonparametric estimators for the cumulative cause specific reversed hazard rates is obtained as $\hat{H}_j(t) = \sum_{i:t_i \geq t} \hat{h}_j(t_i)$. The asymptotic variance of the estimator can be obtained from the observed information matrix using the technique given in Louis (1982), as mentioned in Dewanji and Sengupta (2003) [15,35].

Weighted least squares estimator

Often, the information about the probability that a particular cause is responsible for the failure from a given set of possible causes is available to us. By incorporating this information, in this subsection, we suggest a nonparametric estimation procedure for cumulative cause specific reversed hazard rates using method of weighted least squares. Suppose that, for each individual, we observe a failure time or censoring time and a set g, representing possible causes of failure.

We define the conditional probability of observing $g \ni j$ as the set of possible causes, given the failure of the component at time t due to cause j as,

$$P_{gj}(t) = P[G = g | T = t, J = j, \delta = 1]. \qquad (3.5)$$

When $j \notin g$, then $P_{gj}(t) = 0$ and for fixed j, $\sum_g P_{gj}(t) = 1$. Assume that the censoring time and missing mechanism are independent. Thus (3.5) can be written as,

$$P_{gj}(t) = P[G = g | T = t, J = j]. \qquad (3.6)$$

We now define the reversed hazard rate for failure due to cause j with $g \ni j$ observed as set of possible causes as,

$$h_{gj}(t) = \lim_{\Delta t \to 0} \left(\frac{P[t - \Delta t \leq T \leq t, G = g, J = j | T \leq t]}{\Delta t} \right) \qquad (3.7)$$

So, that $h_{gj}(t)$ is the product of $h_j(t)$ and $P_{gj}(t)$. Thus (3.7) becomes,

$$h_{gj}(t) = h_{gj}(t) P_{gj}(t). \qquad (3.8)$$

Then the reversed hazard rate at time t with g observed as set of possible causes can be written as,

$$h_g(t) = \sum_{j \in g} h_j(t) P_{gj}(t). \qquad (3.9)$$

Taking summation of (3.9) over all g, we get,

$$h(t) = \sum_g h_g(t). \qquad (3.10)$$

Then the cumulative reversed hazard rate at time t with g observed as set of possible causes is

$$H_g(t) = \int_t^\infty h_g(u) du \qquad (3.11)$$

Note that the probabilities $P_{gj}(t)'$ are to be estimated in practice. In order to estimate these probabilities in practice, we make a convenient assumption that $\hat{P}_{gj}(t)$ is independent of time t and henceforth denote as P_{gj}. Using (3.9) we have

$$H^*_g(t) = P\,H^*(t) \qquad (3.12)$$

Where $H^*_g(t)$ is the $(2^k - 1) \times 1$ vector of Hg (t)'s, $H^*_g(t)$'s, is the $k \times 1$ vector of $H_j(t)$'s and P is the $(2^k - 1) \times k$ matrix of P_{gj}'s.

To estimate $H^*_g(t)$, we observe n independently and identically distributed observations $(T_i, \delta_i, \delta_i g_i)$, $i = 1, 2, ..n$; where $T_i = max(X_i, C_i)$, $\delta_i = I(X_i = T_i)$ and g be the set of possible causes associated with the failure of i th individual. Consider the $(2^k - 1)$ dimensional counting process $\{N_g(t)\} g \in G$, where G contains all the non-empty subsets of $\{1, 2...k\}$ and $N_g(t)$ represents the number of events occurring in (t, τ), with g as the observed set of possible causes. Assume that the point of reference, τ is far away from the time span of interest. Now we define

$$dM_g(t) = dN_g(t) - Y(t)dH_g(t). \qquad (3.13)$$

Following Andersen, *et al.* (2003) we show in Appendix A that $M_g(t)$'s are the local square integrable martingales [36]. Then, the nonparametric estimator of $H_g(t)$ is obtained as,

$$H_g(t) = \int_t^{to} \frac{C(u)dN_g(u)}{Y(u)} \qquad (3.14)$$

Where $Y(t)$ is the number of failures up to time t, $C(u) = I(Y(u) > 0)$ and $t_0 = \inf(t; F(t) < 1)$. The details are given in Appendix A.

Now, using martingale central limit theorem we see that (3.14) converges to a Gaussian process with mean $H_g(t)$ and variance $\sigma^2(t)$, which is consistently estimated by,

$$\sigma^2(t) = \int_t^{to} \frac{C(u)dN_g(u)}{Y^2(u)} \qquad (3.15)$$

Using (3.12) and (3.14) we get,

$$\hat{H}^*_g(t) = PH^*(t) + \in(t) \qquad (3.16)$$

Where $\hat{H}^*_g(t)$ is the vector of $\hat{H}_g(t)$ and $\in(t)$ is a vector of Gaussian martingales whose variance is consistently estimated by the matrix $\mathrm{diag}(\hat{\sigma}^2(t))$. Now (3.16) is in the form of a linear model with P to be estimated. Let $\hat{P}$ be a consistent estimator of P. Then by using the principle of weighted least squares, a consistent estimator of $H^*(t)$ is,

$$\hat{H}^*(t) = ((\hat{P}^T V(t)\hat{P})^{-1} \hat{P}^T V(t)\hat{H}^*_g(t) \qquad (3.17)$$

where $V(t)$ is the inverse of the estimated diagonal covariance matrix of $\hat{H}^*_g(t)$, which is given by $V(t) = diag\left(\dfrac{1}{\hat{\sigma}^2(t)}\right).For\ j \in g,$

$$P_{gi} = P[G = g \,|\, J = j] = \frac{P[J = j \,|G = g]\,P[G = g]}{\sum_{g' \in G} P[J = j \,|\, G = g']\,P[G = g']} \qquad (3.18)$$

where G contains all the non-empty subsets of $\{1, 2, ...k\}$. We denote $f_g = P[G = g]$, and $q_{jg} = P[J = j|G = g]$. Thus P_{gj} can be estimated using (3.18), from the values of f_g and q_{jg}. The information about q_{jg} may be available which can utilized to estimate P_{gj}.

For fixed t, $\hat{H}^*_g(t)$ converges in distribution to a $(2^k - 1)$ variate normal with mean $H^*_g(t)$. Then $\hat{H}^*(t)$ converges in distribution to a k variate normal with mean $H^*(t)$ and variance covariance matrix is estimated consistently by $(\hat{P}^T V(t)\hat{P})^{-1}$. This asymptotic variance can be used for constructing large sample confidence limits.

The nonparametric estimator of $F_j(t)$ is given by

$$\hat{F}_j(t) = -\int_0^t \hat{F}(u)d\hat{H}_j(u) \qquad (3.19)$$

Where

$$\hat{F}(t) = exp[-\hat{H}(t)] = exp\left[\sum_{j=1}^k \hat{H}_j(t)\right] \qquad (3.20)$$

where $\hat{H}(t) = \sum_{j=1}^k \hat{H}_j(t)$ is the estimator of cumulative reversed hazard rate for T. Note that the least squares estimator of $H_j(t)$, may not be monotone and it may violate non increasing nature at some points. To develop the estimator with monotonic decreasing property one can use the pooling-the-adjacent-violators algorithm.

Now consider the testing of null hypothesis, $H_0 : H_j(t) = H_{0j}(t)$ for all j, and a fixed t'.

We can formulate the test statistic as,

$$\chi^2(t') = \left(\hat{H}*(t')\right) T \left(\hat{P}^T V(t') \hat{P}\right) (\hat{H}^*(t') - H_0^*(t')) \quad (3.21)$$

where $H_0^*(t)$ is the vector of $H_{0j}(t)$. Since for fixed t', $\hat{H}*(t')$ converges in distribution to a k variate normal, the distribution of (3.21) is asymptotically chi-square with k degrees of freedom under H_0. Thus we reject H_0 if $\chi^2(t') > \chi^2(\alpha)$.

Simulation study

To asses the performance of the estimators, we carried out a simulation study. Suppose there are two causes of failure. We generate random samples from the following parametric family of sub-distribution functions proposed by Dewan & Kulathinal [37].

Let

$$F_1(t) = \underline{P}\,[T \le t, J = 1] = \phi F^a(t)$$

and

$$F_2(t) = \underline{P}\,[T \le t, J = 2] = F(t) - \phi F^a(t) \quad (4.1)$$

where $1 \le a \le 2$, $0 \le \phi \le 0.5$ and $F(t)$ is the distribution function of failure time T. Note that $\phi = P\,[J = 1]$. The restriction on the parameters are imposed due to nonnegativity condition of cause specific density function of T. From (4.1), we obtain $h_1(t) = \phi a F^{a-2}(t) f(t)$ and $h_2(t) = h(t) - h_1(t)$.

Let the failure time distribution be exponential with $F(t) = 1 - exp[-\lambda t]$. Censored observations are generated from $U(0, b)$ where b is chosen such a way that approximately 20% of the observations are left censored. We generate random sample of sizes $n = 100$, and 250. The masked set $g = \{1, 2\}$ is randomly allocated to the observed lifetimes so that the chance for an observed lifetime to be masked is 0.5. We evaluated the estimates of $H_1(t)$ and $H_2(t)$ using two different methods in Nonparametric estimation.

We first computed the maximum likelihood estimates of the $H_j(t)$, $j = 1, 2$. Based on 1000 simulation, we compute absolute bias and mean squared error (MSE) of the estimates for different parametric values of λ, a and ϕ. Simulation study shows that the bias and MSE of the estimates do not vary much with different values of λ, a and ϕ. We therefore present results for two parametric combinations of λ, a and ϕ, which are given in Tables 1 and 2.

n	t	20% censored				Un censored			
		$\widehat{H}_1(t)$		$\widehat{H}_2(t)$		$\widehat{H}_1(t)$		$\widehat{H}_2(t)$	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
100	0.5	0.1273	0.0413	0.1293	0.0700	0.1254	0.0481	0.1287	0.0619
	1	0.1214	0.0342	0.1201	0.0500	0.1038	0.0210	0.1194	0.0211
	1.5	0.0973	0.0024	0.0654	0.0768	0.0704	0.0031	0.0528	0.0041
	2	0.0554	0.0091	0.0324	0.0413	0.0364	0.0021	0.0252	0.0056
250	0.5	0.1179	0.0220	0.0932	0.0644	0.1151	0.0223	0.0513	0.0120
	1	0.0984	0.0090	0.0511	0.0321	0.0807	0.0103	0.0429	0.0070
	1.5	0.0532	0.0010	0.0217	0.0520	0.0451	0.0021	0.0346	0.0024
	2	0.0333	0.0030	0.0113	0.0414	0.0138	0.0002	0.0093	0.0009

Table 1: Bias and MSE of the maximum likelihood estimates of $H_j(t)$, $j = 1, 2$ for $\lambda = 0.5$, $a = 1.5$, and $\phi = 0.4$

We then computed weighted least squares estimates $\hat{H}_1(t)$ and $\hat{H}_2(t)$ at different time points and for three cases of q_{jg}. The choices of q_{jg}'s for the masked set $g = \{1, 2\}$ are,

(i) $q_{1g} = 0.5$ and $q_{2g} = 0.5$ (equal probability for cause 1 and cause 2)

(ii) $q_{1g} = 0.992$ and $q_{2g} = 0.008$ (greater probability for cause 1)

and (iii) $q_{1g} = 0.008$ and $q_{2g} = 0.992$ (greater probability for cause 2)

These q_{jg}'s are used to estimate P_{gj}'s. Based on 1000 replications, we computed absolute bias and MSE of the estimates of $H_j(t)$, $j = 1$, 2. Tables 3-5 provide bias and MSE of least squares estimates for different sample sizes.

n	t	20% censored				Un censored			
		$\widehat{\widehat{H}}_1(t)$		$\widehat{\widehat{H}}_2(t)$		$\widehat{\widehat{H}}_1(t)$		$\widehat{\widehat{H}}_2(t)$	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
100	0.5	0.1143	0.0753	0.0932	0.0082	0.0994	0.0153	0.0704	0.0053
	1	0.0932	0.0803	0.0817	0.0073	0.0637	0.0040	0.0468	0.0020
	1.5	0.0734	0.0043	0.0532	0.0052	0.0515	0.0048	0.0393	0.0050
	2	0.0532	0.0094	0.0316	0.0013	0.0344	0.0029	0.0229	0.0024
250	0.5	0.0982	0.0532	0.0913	0.0090	0.0663	0.0045	0.0636	0.0050
	1	0.0734	0.0093	0.0432	0.0034	0.0538	0.0032	0.0317	0.0022
	1.5	0.0666	0.0031	0.0224	0.0021	0.0471	0.0035	0.0171	0.0031
	2	0.0326	0.0060	0.0210	0.0011	0.0288	0.0006	0.0092	0.0001

Table 2: Bias and MSE of the maximum likelihood estimates of $H_j(t)$, $j = 1, 2$ for $\lambda = 1$, $\alpha = 1.75$, and $\phi = 0.2$

Simulation study shows that both bias and MSE decreases as sample size increases and slightly increases as censoring percentage increases. Moreover, as the time increases bias reduces. It may due to the fact that the left tail of observations are more affected by the censored observations. It may be noted that in the case of least squares estimator, bias is small and pretty close to each other for $\widehat{H}_1(t)$ and $\widehat{H}_2(t)$ when q_{jg}'s are equal.

When probability of occurrence of failure due to cause 1 is larger than that of cause 2 ($q_{1g} > q_{2g}$), then the bias is slightly less for $\widehat{H}_1(t)$ compared to $\widehat{H}_1(t)$, as expected. Similarly, bias shows lesser values for $\widehat{H}_2(t)$ if probability of occurrence of failure due to cause 2 is larger than that of cause 1 ($q_{2g} > q_{1g}$).

	n	t	20% censored				Un censored			
			$\widehat{\widehat{H}}_1(t)$		$\widehat{\widehat{H}}_2(t)$		$\widehat{\widehat{H}}_1(t)$		$\widehat{\widehat{H}}_2(t)$	
			Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
a=1.75 ϕ=0.2 λ=1	100	1	0.0574	0.0039	0.0372	0.0005	0.0482	0.0036	0.0213	0.0005
		1.5	0.0319	0.0014	0.0212	0.0009	0.0213	0.0011	0.0200	0.0004
		2	0.0174	0.0009	0.0137	0.0005	0.0124	0.0008	0.0113	0.0003
		2.5	0.0077	0.0002	0.0087	0.0002	0.0073	0.0001	0.0023	0.0001
	250	1	0.0572	0.0035	0.0365	0.0003	0.0437	0.0023	0.0210	0.0070
		1.5	0.0313	0.0012	0.0201	0.0007	0.0223	0.0011	0.0117	0.006
		2	0.0163	0.0005	0.0128	0.0005	0.0103	0.0004	0.0032	0.0002
		2.5	0.0064	0.0002	0.0087	0.0001	0.0050	0.0002	0.0017	0.0001
a=1.5 ϕ=0.4 λ=0.5	100	1	0.1613	0.0851	0.1787	0.0897	0.0732	0.0532	0.0706	0.0327
		1.5	0.0938	0.0047	0.0940	0.0054	0.0703	0.0086	0.0632	0.0034
		2	0.0666	0.0021	0.0592	0.0036	0.0584	0.0013	0.0319	0.0021
		2.5	0.0162	0.0112	0.0179	0.00191	0.0111	0.0010	0.0133	0.0020
	250	1	0.1591	0.0844	0.1641	0.0888	0.0687	0.03742	0.0538	0.0310
		1.5	0.0811	0.0042	0.0729	0.0051	0.0573	0.0032	0.0432	0.0030
		2	0.0655	0.0022	0.0505	0.0032	0.0329	0.0023	0.0211	0.0025
		2.5	0.0260	0.0011	0.0175	0.0019	0.0113	0.0010	0.0120	0.0013

Table 3: Bias and MSE of least squares estimates of $H_j(t)$, $j = 1, 2$ for $q_{1\{12\}} = 0.5$ and $q_{2\{12\}} = 0.5$

	n	t	20% censored $\widehat{\widehat{H}}_1(t)$ Bias MSE	$\widehat{\widehat{H}}_2(t)$ Bias MSE	Un censored $\widehat{\widehat{H}}_1(t)$ Bias MSE	$\widehat{\widehat{H}}_2(t)$ Bias MSE
$a=1.75$ $\phi=0.2$ $\lambda=1$	100	1	0.1002 0.0106	0.0857 0.0095	0.0932 0.0096	0.0782 0.0073
		1.5	0.0562 0.0035	0.0499 0.0038	0.0326 0.0034	0.0336 0.0032
		2	0.0326 0.0012	0.0310 0.0018	0.0118 0.0022	0.0211 0.0023
		2.5	0.0188 0.0005	0.0209 0.0009	0.0093 0.0010	0.0101 0.0009
	250	1	0.0995 0.0104	0.0754 0.0063	0.0812 0.0093	0.0513 0.0032
		1.5	0.0555 0.0033	0.0433 0.0023	0.0232 0.0021	0.0232 0.0024
		2	0.0309 0.0012	0.0259 0.0009	0.0116 0.0010	0.0101 0.0017
		2.5	0.0171 0.0004	0.0155 0.0004	0.0090 0.0006	0.0089 0.0015
$a=1.5$ $\phi=0.4$ $\lambda=0.5$	100	1	0.1698 0.0888	0.1576 0.0832	0.1113 0.0773	0.0993 0.0572
		1.5	0.1482 0.0439	0.1041 0.0513	0.0932 0.0349	0.0739 0.0474
		2	0.0982 0.0072	0.0862 0.0071	0.0715 0.0027	0.0632 0.0032
		2.5	0.0307 0.0052	0.0143 0.0021	0.0232 0.0010	0.0114 0.0021
	250	1	0.1602 0.0885	0.1492 0.0826	0.0913 0.0432	0.0732 0.0513
		1.5	0.1414 0.0431	0.0996 0.0092	0.0737 0.0132	0.0651 0.0428
		2	0.0835 0.0022	0.0764 0.0051	0.0632 0.0017	0.0318 0.0024
		2.5	0.0211 0.0016	0.0111 0.0018	0.0187 0.0011	0.0099 0.0011

Table 4: Bias and MSE of least squares estimates of $H_j(t)$, $j = 1, 2$ for $q_{1\{12\}} = 0.008$ and $q_{2\{12\}} = 0.992$

	n	t	20% censored $\widehat{\widehat{H}}_1(t)$ Bias MSE	$\widehat{\widehat{H}}_2(t)$ Bias MSE	Un censored $\widehat{\widehat{H}}_1(t)$ Bias MSE	$\widehat{\widehat{H}}_2(t)$ Bias MSE
$a=1.75$ $\phi=0.2$ $\lambda=1$	100	1	0.0270 0.0026	0.0456 0.0031	0.0232 0.0017	0.0356 0.0023
		1.5	0.0186 0.0027	0.0271 0.0009	0.0117 0.0010	0.0264 0.0021
		2	0.0135 0.0005	0.0259 0.0004	0.0108 0.0003	0.0201 0.0009
		2.5	0.0094 0.0002	0.0092 0.0001	0.0082 0.0002	0.0052 0.0005
	250	1	0.0216 0.0011	0.0447 0.0029	0.0210 0.0020	0.0432 0.0032
		1.5	0.0142 0.0021	0.0264 0.0008	0.0193 0.0032	0.0236 0.0007
		2	0.0092 0.0001	0.0157 0.0003	0.0072 0.0021	0.0063 0.0001
		2.5	0.0065 0.0003	0.0088 0.0037	0.0037 0.00030	0.0063 0.0001
$a=1.5$ $\phi=0.4$ $\lambda=0.5$	100	1	0.1326 0.0728	0.1451 0.0316	0.1013 0.0376	0.1358 0.0216
		1.5	0.0973 0.0191	0.1241 0.0577	0.0672 0.0413	0.1143 0.0132
		2	0.0144 0.0024	0.0829 0.0033	0.0133 0.0017	0.0771 0.0032
		2.5	0.0081 0.0001	0.0195 0.0031	0.0072 0.00090	0.0114 0.0014
	250	1	0.1301 0.0701	0.1440 0.0303	0.0932 0.0326	0.0310 0.0203
		1.5	0.0961 0.0182	0.1238 0.0570	0.0432 0.0104	0.1124 0.0100
		2	0.0141 0.0021	0.0715 0.0032	0.0123 0.0010	0.0632 0.0030
		2.5	0.0077 0.0001	0.0185 0.0013	0.0050 0.00030	0.0104 0.0010

Table 5: Bias and MSE of least squares estimates of $H_j(t)$, $j = 1, 2$ for $q_{1\{12\}} = 0.992$ and $q_{2\{12\}} = 0.008$

Data analysis

To illustrate the proposed methodology, we consider the hard drives data given in Flehinger, *et al.* (1998) [13]. The data provide the failure times of hard drives of computer and corresponding cause of failures. There are 3 causes of failure denoted as 1, 2 and 3. We assume these causes act independently. All together 172 failures are reported in the study period, in which some of the failures are masked. The data are obtained from a two stage experimental procedure and we combine the data of two stages and apply the methods given in Nonparametric estimation to the data. Exact cause of failure is available for 66 out of 172 hard drives. The only observed masked group were {1, 2, 3} and {1, 3}.

The data set presented in first four columns of table given in Appendix A of Flehinger, *et al.* (1998) [13]. In table, second column represents the failure time. The third column (outcome) gives the cause of failure if it is identified correctly or resolved in Stage 2. Here -1 in third column indicates the unresolved problems. Fourth column gives the information about masking. We make the 20% of data randomly to be left censored. We first consider Maximum likelihood estimator method and computed the maximum likelihood estimates of $H_j(t)$ j = 1, 2, 3.

In order to illustrate the Weighted least squares estimator method for calculating the least squares estimates of $H_j(t)$ j = 1, 2, 3, we require the probabilities P_{gj}. We consider $P_{\{1,3\}1}$ = 0.412, $P_{\{1,3\}3}$ = 0.446 and $P_{\{1,2,3\}1}$ = 0.310, $P_{\{1,2,3\}2}$ = 0.469, $P_{\{1,2,3\}2}$ = 0.436. These probabilities are estimated in Flehinger, *et al.* by assuming Weibull model for the data [13]. Now we compute the least squares estimates of $H_j(t)$ j = 1, 2, 3. We use pooling-the-adjacent-violators algorithm to achieve the monotonicity of the least squares estimates. Figure 1 depicts the plots of maximum likelihood estimates and least squares estimates of $H_j(t)$ j = 1, 2, 3 along with 95% confidence limits.
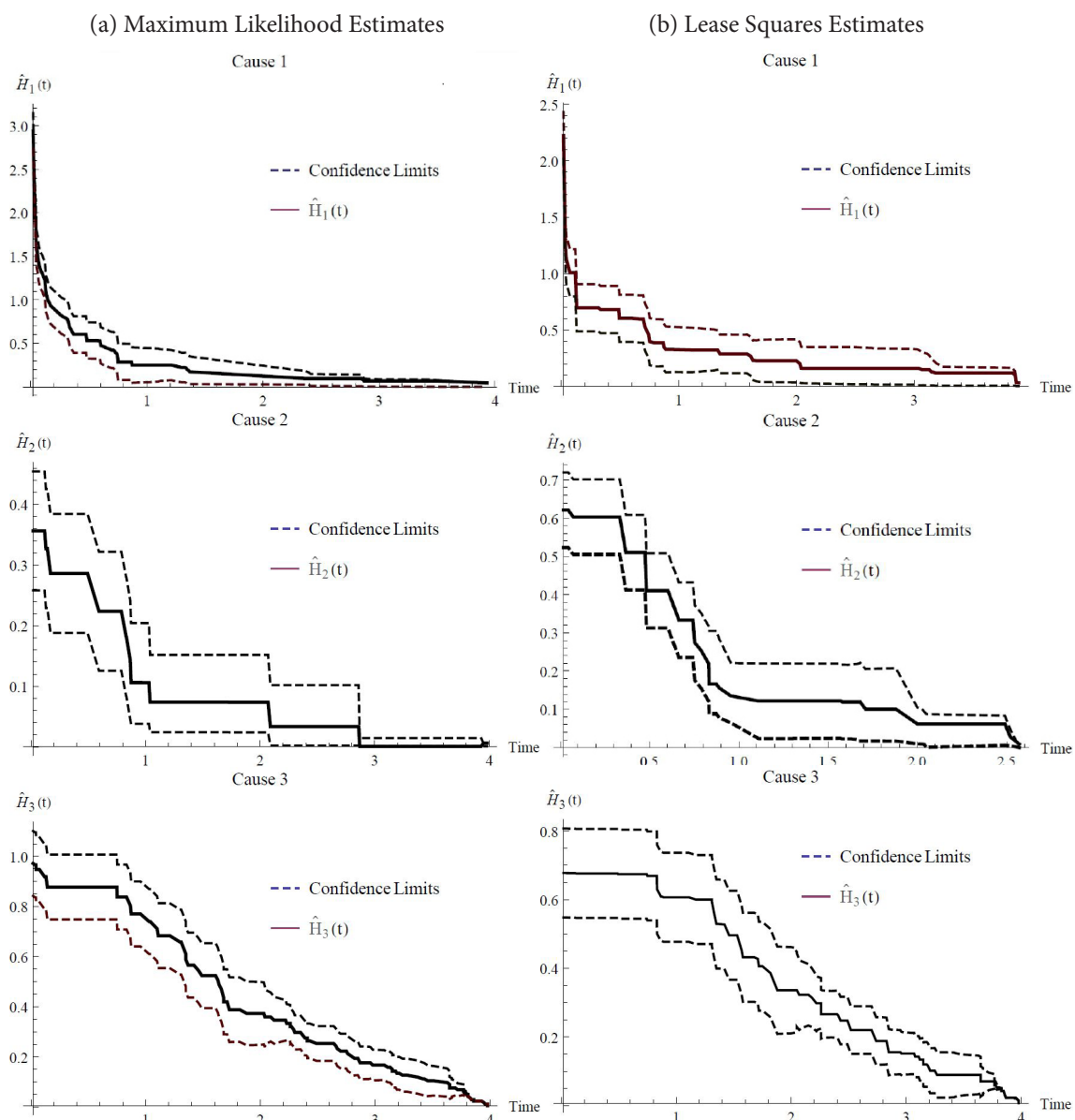

Figure 1: Plot of **(a)** maximum likelihood estimates and **(b)** least square estimates of $H_j(t)$ j = 1, 2, 3

The plots of estimates of cumulative incidence functions for three different causes are given in Figures 2 and 3. Figures 2 and 3 show that, at early stages, the majority of the failure is due to cause 1 and after a certain time period (around $t=2.7$), failure due to cause 3 dominates.
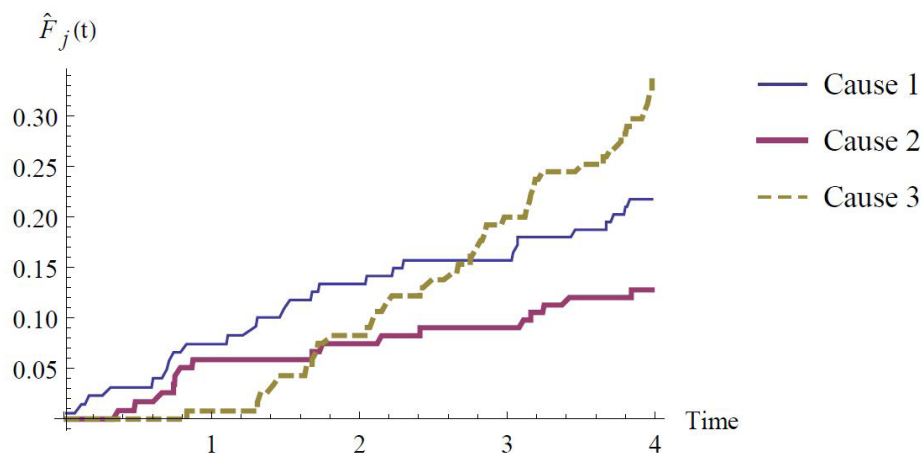
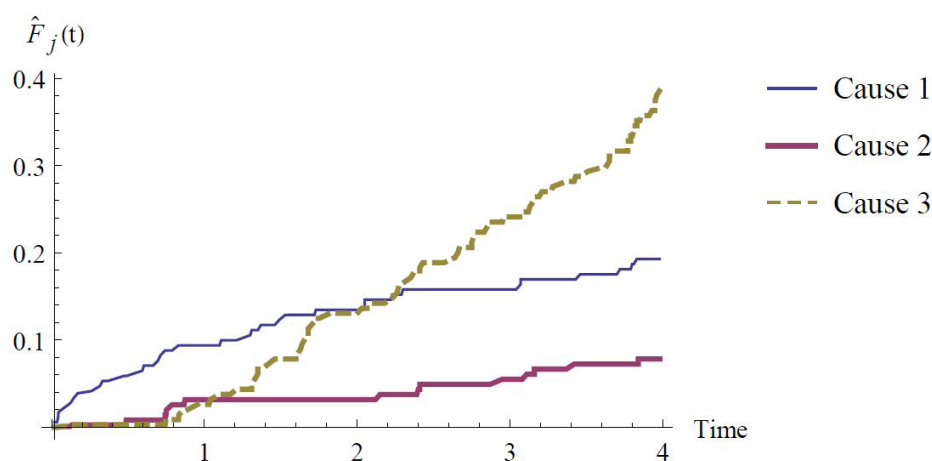$\hat{F}_j(t)$

Figure 2: Plot of maximum likelihood estimate of $F_j(t)$

$\hat{F}_j(t)$

Figure 3: Plot of least squares estimate of $F_j(t)$

Conclusion

References

1. Crowder MJ (2001) Classical Competing Risks. CRC Press, London.

2. Kalbfleisch JD, Prentice RL (2002) The Statistical Analysis of Failure Time Data. John Wiley & Sons New York.

3. Lawless JF (2003) Statistical Models and Methods for Lifetime Data. John Wiley & Sons New York.

4. Dinse GE (1982) Nonparametric estimation for partially-complete time and type of failure data. Biometrics 38: 417-31.

5. Miyakawa M (1984) Analysis of incomplete data in competing risks model. IEEE Transactions on Reliability 33: 293-96.

6. Racine-Poon AH, Hoel DG (1984) Nonparametric estimation of the survival function when cause of death is uncertain. Biometrics 40: 1151-8.

7. Lo SH (1991) Estimating a survival function with incomplete cause-of-death data. J Multivariate Anal 39: 217-35.

8. Mukerjee H, Wang JL (1993) Nonparametric maximum likelihood estimation of an increasing hazard rate for uncertain cause-of-death data. Scandinavian J Stat 17-33.

9. Goetghebeur E, Ryan L (1990) A modified log rank test for competing risks with missing failure type. Biometrika 77: 207-11.

10. Goetghebeur E, Ryan L (1995) Analysis of competing risks survival data when some failure types are missing. Biometrika 82: 821-33.

11. Dewanji A (1992) A note on a test for competing risks with missing failure type. Biometrika 79: 855-57.

12. Lu K, Tsiatis AA (2001) Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. Biometrics 57: 1191-7.

13. Flehinger BJ, Reiser B, Yashchin E (1998) Survival with competing risks and masked causes of failures. Biometrika 85: 151-64.

14. Flehinger BJ, Reiser B, Yashchin E (2002) Parametric modeling for survival with competing risks and masked failure causes. Lifetime Data Anal 8: 177-203.

15. Dewanji A, Sengupta D (2003) Estimation of competing risks with general missing pattern in failure types. Biometrics 59: 1063-70.

16. Sen A, Banerjee M, Li Y, Noone AM (2010) A Bayesian approach to competing risks analysis with masked cause of death. Stat Med 29: 1681-95.

17. Hyun S, Lee J, Sun Y (2012) Proportional hazards model for competing risks data with missing cause of failure. J Stat Planning Inf 142: 1767-79.

18. Craiu RV, Duchesne T (2004) Inference based on the em algorithm for the competing risks model with masked causes of failure. Biometrika 91: 543-58.

19. Craiu RV, Reiser B (2006) Inference for the dependent competing risks model with masked causes of failure. Lifetime Data Anal 12: 21-33.

20. Lee M1, Dignam JJ, Han J (2014) Multiple imputation methods for inference on cumulative incidence with missing cause of failure. Stat Med 33: 4605-26.

21. Barlow RE, Marshall AW, Proschan F (1963) Properties of probability distributions with monotone hazard rate. Annals Math Stat 34: 375-89.

22. Wagner SS, Altmann SA (1973) What time do the baboons come down from the trees?(An estimation problem). Biometrics 29: 623-35.

23. Ware JH, Demets DL (1976) Reanalysis of some baboon descent data. Biometrics 32: 459-63.

24. Keilson J, Sumita U (1982) Uniform stochastic ordering and related inequalities. Can J Stat 10: 181-98.

25. Block HW, Savits TH, Singh H (1998) The reversed hazard rate function. Probability Engineer Info Sci, 12 (1), 69-90.

26. Finkelstein MS (2002) On the reversed hazard rate. Reliability Engineer Syst Safety 78: 71-5.

27. Nair NU, Sankaran PG, Asha G (2005) Characterizations based on reliability concepts. J Appl Stat Sci 14: 237-42.

28. Sankaran PG, Gleeja VL (2011) On proportional reversed hazards frailty models. Metron LXIX: 151-73.

29. Sankaran PG, Anjana S (2014) A class of tests for the equality of cause- specific reversed hazard rates in competing risks models. J Indian Stat Assoc 52: 161-76.

30. Sankaran P, Anjana S (2015) Proportional cause-specific reversed hazards model. J Nonparametric Stat 1-16.

31. Dewanji A, Sankaran P, Sengupta D, Karmakar B (2016) Regression anal- ysis of competing risks data with general missing pattern in failure types. Stat Methodol 29: 18-31.

32. Little RJA, Rubin DB (1987) Statistical Analysis with Missing Data. John Wiley & Sons, New York.

33. Dempster AP, Laird NM, Rubin, DB (1977) Maximum likelihood from incomplete data via the em algorithm. J Royal Stat Soci 1-38.

34. Wu CF (1983) On the convergence properties of the EM algorithm. Annals Stat 11: 95-103.

35. Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J Royal Stat Soci B44: 226-33.

36. Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical Models Based on Counting Processes. Springer-Verlag, New York.

37. Dewan I, Kulathinal S (2007) On testing dependence between time to failure and cause of failure when causes of failure are missing. PloS One 2: e1255.