Simulation Study under a Semi-parametric Model for Censored Gap Times

Amorim, A.P.[1], Moreira, C.[2]

[1] University of Minho-Centre of Mathematics, Portugal
[2] University of Minho-Centre of Mathematics, Portugal

E-mail for correspondence: apamorim@math.uminho.pt

Abstract: We consider a comparison between the Kaplan-Meier and the semi-parametric estimators for a censorship models. The observations are assumed to be generated under a semi-parametric random censorship, this mean that a random censorship model where de conditional expectation of censoring indicator given the observations belongs to a parametric family. The semi-parametric estimator of the survival function was defined in de Uña-Alvarez and Amorim (2011). An asymptotic representation of a general empirical integral as a sum of independent and identically distributed (i.i.d.) random variables under the proposed model was obtained in Amorim (2012). The performance of the corresponding asymptotic confidence intervals (a.c.i.) relative to that of a nonparametric method, de Uña-Alvarez and Meira-Machado (2008), is investigated through simulations Dikta et al. (2005).

Keywords: Asymptotic normality, Gap times, Semi-parametric censorship model, Kaplan-Meier, Presmoothing.

1 Introduction

Let (T_1, T_2) be a pair of gap times of successive events, which are observed subject to random right-censoring. Let C be the right-censoring variable, assumed to be independent of (T_1, T_2), and let $Y = T_1 + T_2$ be the total time. Due to censoring, rather than (T_1, T_2) we observe $\left(\widetilde{T}_1, \widetilde{T}_2, \Delta_1, \Delta_2\right)$, where $\widetilde{T}_1 = T_1 \wedge C$, $\Delta_1 = I(T_1 \leq C)$ and $\widetilde{T}_2 = T_2 \wedge C_2$, $\Delta_2 = I(T_2 \leq C_2)$, where $C_2 = (C - T_1)\, I\,(T_1 \leq C)$ is the censoring variable for the second gap time. When $\Delta_2 = 1$ then $\Delta_1 = 1$. Hence, $\Delta_2 = \Delta_1\Delta_2 = I\,(Y \leq C)$ is the censoring indicator pertaining to the total time. We put $\widetilde{Y} = Y \wedge C$.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), Guimarães, Portugal, 7–12 July 2019.

Let $\left(\widetilde{T}_{1i}, \widetilde{T}_{2i}, \Delta_{1i}, \Delta_{2i}\right)$, $1 \leq i \leq n$, be iid data with the same distribution as $\left(\widetilde{T}_1, \widetilde{T}_2, \Delta_1, \Delta_2\right)$. The censoring time C is assumed to be independent of the pair (T_1, T_2). The marginal distribution of the first gap time T_1 may be consistently estimated by the Kaplan-Meier estimator based on the $\left(\widetilde{T}_{1i}, \Delta_{1i}\right)$'s. Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on the $\left(\widetilde{T}_{1i} + \widetilde{T}_{2i}, \Delta_{2i}\right)$'s. However, T_2 and C_2 will be in general dependent (because the expected correlation between the gap times), and hence the estimation of the marginal distribution of the second gap time is not such a simple issue. Also, it is not clear in principle how the bivariate distribution function $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$ can be efficiently estimated. This issue was investigated, among others, by Lin et al. (1999), Schaubel and Cai (2004), or de Uña-Alvarez and Meira-Machado (2008).

2 Estimation of bivariate distribution function

De Uña-Alvarez and Amorim (2011) proposed a semiparametric estimator for the bivariate distribution function of the gap times, $F_{12}(x, y)$ by presmoothing the estimator of de Uña-Alvarez and Meira-Machado (2008). The probability of censoring for T_2 given the (possibly censored) gap times belongs to a parametric family of binary regression curves; $m(x, y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y)$, where $m(x, y)$ follows some parametric model, $m(\cdot, \cdot; \beta)$, and the parametric model for m is estimated from the observable data. The censoring indicators Δ_{2i}'s is replaced by the fitted values of m. Some notations must be considered: $\widetilde{Y}_i = \widetilde{T}_{1i} + \widetilde{T}_{2i}$ be the $i-$th recorded total time; the ordered $\widetilde{Y}-$statistics $\widetilde{Y}_{1:n} \leq \widetilde{Y}_{2:n} \leq \ldots \leq \widetilde{Y}_{n:n}$; and the $(\widetilde{T}_{[1i:n]}, \widetilde{T}_{[2i:n]}, \Delta_{[1i:n]}, \Delta_{[2i:n]})$ the $i-$th concomitant, i.e. the information attached to $\widetilde{Y}_{i:n}$.
Let W_i be the Kaplan-Meier weight attached to $\widetilde{Y}_{i:n}$

$$W_i = \frac{\Delta_{[2i:n]}}{n - i + 1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{[2j:n]}}{n - j + 1}\right].$$

The estimator in de Uña-Alvarez and Meira-Machado (2008) is:

$$\widehat{F}_{12}(x, y) = \sum_{i=1}^{n} W_i I(\widetilde{T}_{[1i:n]} \leq x, \widetilde{T}_{[2i:n]} \leq y)$$

and the

$$\lim_{n \to \infty} \widehat{F}_{12}(x, y) = P\left(T_1 \leq x, T_2 \leq y, T_1 + T_2 \leq \tau_H\right) \equiv F_{12}^0(x, y)$$

where τ_H is the upper bound of the support of the distribution function H of $\widetilde{Y}$.
Parametrical presmoothed Kaplan-Meier weights are defined as

$$W_i(\beta_n) = \frac{m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}; \beta_n)}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{[1j:n]}, \widetilde{Y}_{j:n}; \beta_n)}{n-j+1}\right]$$

and

$$\widehat{F}_{12}^{sp}(x,y) = \sum_{i=1}^{n} W_i(\beta_n) I(\widetilde{T}_{[1i:n]} \leq x, \widetilde{T}_{[2i:n]} \leq y)$$

where β_n is the maximizer of the conditional likelihood

$$L_1(\beta) = \prod_{\Delta_{1i}=1} m\left(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta\right)^{\Delta_{2i}} \left[1 - m\left(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta\right)\right]^{1-\Delta_{2i}},$$

By noting $S(\varphi) = \int \varphi dF_{12}$, we introduce the following estimator of this expectation:

$$S_n(\varphi) = \int \varphi d\widehat{F}_{12}^{sp} = \sum_{i=1}^{n} W_i(\beta_n) \varphi(\widetilde{T}_{[1i:n]}, \widetilde{T}_{[2i:n]}).$$

The $\widehat{F}_{12}^{sp}(x,y)$ is obtained when $\varphi(u,v) = I(u \leq x, v \leq y)$.
The asymptotic representation of $S_n(\varphi)$ as a sum of i.i.d. random variables is in Amorim (2011). The result is similar to those obtained by Stute (1995) for Kaplan-Meier integrals and Dikta (2005) for presmoothed Kaplan-Meier integrals. From the i.i.d. representation and the Central Theorem Limit, the asymptotic normality of $S_n(\varphi)$ is obtained by adaptation of Dikta (2005) to the bivariate setting.

3 Simulation Study

We consider the simple bootstrap as the method to approximate the distribution of the $\widehat{F}_{12}^{sp}$ and $\widehat{F}_{12}$ in finite samples. In our simulations below, 95% confidence intervals are calculated by the mean and the standard deviation of values obtained in simulations for both estimators. The simulated scenario is the same as that described in Lin et al. (1999) and de Uña-Alvarez and Meira-Machado (2008). To be precise, the gap times (T_1, T_2) were generated according to the bivariate distribution

$$F_{12}(x,y) = F_1(x)F_2(y)\left[1 + \theta\left\{1 - F_1(x)\right\}\left\{1 - F_2(y)\right\}\right]$$

where the marginal distribution functions F_1 and F_2 are exponential with rate parameter 1. The single parameter θ controls the amount of dependency between the gap times. $\theta = 0$ for simulating independent gap times; $\theta = 1$ corresponding to 0.25 correlation between T_1 and T_2. An independent uniform censoring time C was generated, according to models:

- Unif$[0, 4]$ (about 24% and 47% of censoring on T_1 and of censoring on T_2);
- Unif$[0, 3]$ (about 32% and 57% of censoring on T_1 and of censoring on T_2);

Sample sizes of n=50, n=100 and n=200 were considered. The number of bootstrap resamples was taken to be B=100. We performed M=1,000 trials for each situations. Results of coverage and 95% confidence intervals are displayed in Table 1 and Table 2. Four different points, of pairs (x, y) corresponding to the four different combinations of the percentiles 20% and 80% of the marginal distributions of the gap times are considered. We calculate both estimators $\widehat{F}_{12}^{sp}(x, y)$ and $\widehat{F}_{12}^{km}(x, y)$ along the 1,000 trials. For $x < y$

$$m(x, y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y) = \frac{1}{1 + \eta(x, y)},$$

where

$$\eta(x, y) = \frac{\lambda_G(y)}{\lambda_{2|1}(y - x|x)},$$

where $\lambda_G(.)$ and $\lambda_{2|1}(.|x)$ stand for the hazard rate functions of C and T_2 given $T_1 = x$, respectively.
$\lambda_G(y) = 1/(\tau_G - y)$ when $C \sim U\,[0, \tau_G]$ and that $\lambda_{2|1}(.|x)$ is given by

$$\lambda_{2|1}(y - x|x) = \frac{2 + 4\exp(-y) - 2\exp(-x) - 2\exp(-y + x)}{2 + 2\exp(-y) - 2\exp(-x) - \exp(-y + x)} \quad if\ \theta = 1,$$

being 1 when $\theta = 0$. We obtained a correctly specified model through a preliminary transformation $\eta(T_1, Y)$ of the data (with $\beta = 1$),

$$m(x, y) = \frac{1}{1 + \exp\{\beta \ln \eta(x, y)\}}.$$

4 Conclusion

A semiparametric estimator $\widehat{F}_{12}^{sp}(x, y)$ of the bivariate distribution function of gap times which are observed under censoring has been revisited. The semiparametric estimator is based on a parametric specification (e.g. logistic) of the conditional probability of censoring for the second gap time T_2, given the available information. The performance of the semiparametric estimator is evaluated in a simulation plan with different proportions of censoring. It can be seen that the coverages provided by the semiparametric estimator as well for its competitor estimator are above the nominal 95%, although this problem is somehow mitigated with an increasing sample size. In general, the coverages of the semiparametric estimator outperforms its competitor and the magnitude of the average intervals based on the semiparametric are smaller. As expected, in both estimators, the magnitude of the average intervals deceases as the sample size increase.

TABLE 1. Simulation study of 95% a.c.i. for $F_{12}(x,y)$ along 1,000 simulated samples, case $\theta = 0$. From top to bottom: $(x,y) = \left(F_1^{-1}(0.2), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.8), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.2), F_2^{-1}(0.8)\right)$, and $\left(F_1^{-1}(0.8), F_2^{-1}(0.8)\right)$.

C	$U\,[0,3]$					
n	50	50	100	100	200	200
	Average Length	Coverage	Average Length	Coverage	Average Length	Coverage
$m(.;\beta)$	0.09395	0.882	0.05684	0.920	0.04920	0.926
KM	0.09925	0.845	0.06325	0.910	0.05500	0.919
$m(.;\beta)$	0.21635	0.925	0.12251	0.940	0.10484	0.934
KM	0.23448	0.922	0.13750	0.926	0.11822	0.930
$m(.;\beta)$	0.21301	0.900	0.12284	0.934	0.10570	0.932
KM	0.23163	0.892	0.13774	0.930	0.11894	0.931
$m(.;\beta)$	0.38512	0.936	0.22337	0.950	0.19179	0.948
KM	0.41482	0.927	0.25460	0.942	0.21988	0.939
C	$U\,[0,4]$					
n	50	50	100	100	200	200
	Average Length	Coverage	Average Length	Coverage	Average Length	Coverage
$m(.;\beta)$	0.09328	0.883	0.07015	0.900	0.05010	0.931
KM	0.09835	0.850	0.07692	0.899	0.05513	0.929
$m(.;\beta)$	0.20783	0.920	0.14565	0.930	0.10315	0.944
KM	0.22327	0.926	0.15862	0.934	0.11314	0.946
$m(.;\beta)$	0.20616	0.903	0.14561	0.924	0.10304	0.947
KM	0.22093	0.903	0.15870	0.919	0.11295	0.948
$m(.;\beta)$	0.33135	0.941	0.23198	0.944	0.16251	0.950
KM	0.35410	0.938	0.25030	0.943	0.17647	0.934

Acknowledgments: This research was financed by FCT - Fundação para a Ciência e a Tecnologia, within the Project UID/MAT/00013/2013.

References

Amorim, A.P. (2012). Semiparametric estimation in the non-Markov three-state and illness-death progressive models. *Ph.D. Thesis*. University of Vigo.

de Uña-Alvarez, J. and Amorim, A.P. (2011). A semiparametric estimator of the bivariate distribution function for censored gap times. *Biometrical Journal* , **53**, 113 – 127.

de Uña-Alvarez, J. and Meira-Machado, L. (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters*, **78**, 2440 – 2445.

Dikta, G. and Ghorai, J. and Schmidt, C. (2005). The central limit theorem under semiparametric random censorship models. *Journal of Statistical Planning and Inference*, **127**, 23 – 51.

TABLE 2. Simulation study of 95% a.c.i. for $F_{12}(x,y)$ along 1,000 simulated samples, case $\theta = 1$. From top to bottom: $(x,y) = \left(F_1^{-1}(0.2), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.8), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.2), F_2^{-1}(0.8)\right)$, and $\left(F_1^{-1}(0.8), F_2^{-1}(0.8)\right)$.

C	$U[0,3]$					
n	50	50	100	100	200	200
	Average Length	Coverage	Average Length	Coverage	Average Length	Coverage
$m(.;\beta)$	0.12752	0.871	0.09210	0.921	0.06547	0.939
KM	0.13460	0.870	0.09830	0.930	0.07003	0.929
$m(.;\beta)$	0.21540	0.931	0.15331	0.939	0.10830	0.952
KM	0.23965	0.914	0.17184	0.933	0.12278	0.951
$m(.;\beta)$	0.22544	0.918	0.15974	0.941	0.11312	0.938
KM	0.24054	0.920	0.17197	0.937	0.12190	0.939
$m(.;\beta)$	0.36553	0.935	0.26479	0.940	0.18913	0.951
KM	0.40661	0.910	0.30383	0.935	0.22051	0.939
C	$U[0,4]$					
n	50	50	100	100	200	200
	Average Length	Coverage	Average Length	Coverage	Average Length	Coverage
$m(.;\beta)$	0.12620	0.871	0.09233	0.912	0.06627	0.933
KM	0.13172	0.856	0.09753	0.908	0.06997	0.934
$m(.;\beta)$	0.21414	0.933	0.15042	0.940	0.10690	0.954
KM	0.23216	0.039	0.16432	0.927	0.11705	0.941
$m(.;\beta)$	0.22013	0.906	0.15665	0.927	0.11083	0.951
KM	0.23031	0.904	0.16529	0.933	0.11680	0.944
$m(.;\beta)$	0.31972	0.938	0.22596	0.946	0.15935	0.956
KM	0.34833	0.927	0.24695	0.944	0.17398	0.954

Lin, D. Y. and Sun, W. and Ying, Z. (1999). Nonparametric Estimation of the Gap Time Distributions for Serial Events with Censored Data. *Biometrika* , **86**, 59 – 70.

Schaubel, D. E. and Cai, J. (2004). Non-parametric Estimation of Gap-time Survival Functions for Ordered Multivariate Failure Time Data. *Statistics in Medicine* , **23**, 1885 – 1900.

Stute, W. (1995). The Central Limit Theorem Under Random Censorship. *Scandinavian Journal of Statistics* , **23**, 422 – 439.