Estimation of the bivariate distribution function for censored gap times

Jacobo de Uña-Álvarez[1], Luís Meira-Machado[2]

[1] jacobo@uvigo.es, Departament of Statistics and OR, University of Vigo,

[2] lmachado@mct.uminho.pt, Departament of Mathematics for Science and Technology, University of Minho

Abstract

In many medical studies, patients may experience several events. The times between consecutive events (gap times) are often of interest and lead to problems that have received much attention recently. In this work we consider a new nonparametric estimator of the bivariate distribution function for censored gap times. We explore the behaviour of the estimator through simulations. An illustration through real data analysis is included.

Keywords: bivariate censoring; Kaplan-Meier; nonparametric estimation.

1. Introduction

In longitudinal studies of disease, patients can experience several events through a follow-up period. In these studies, the sequentially ordered events (gap times) are often of interest. The events of concern may be of the same nature (e.g. cancer patients may experience recurrent disease episodes) or represent different states in the disease process (e.g. alive and disease-free, alive with recurrence and dead). If the events are of the same nature this are usually referred as recurrent event, whereas if they represent different states (i.e. multi-state models) they are usually modelled thought their intensity functions [1].

Let (T_1, T_2) be a pair of gap times of successive events, which are observed subjected 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. Because of this, we only observe $\left(\tilde{T}_{1i}, \tilde{T}_{2i}, \Delta_{1i}, \Delta_{2i}\right)$, $1 \le i \le n$, which are n independent replications of $\left(\tilde{T}_1, \tilde{T}_2, \Delta_1, \Delta_2\right)$, where $\tilde{T}_1 = T_1 \wedge C$, $\Delta_1 = \mathbb{I}(T_1 \le C)$, and $\tilde{T}_2 = T_2 \wedge C_2$, $\Delta_2 = \mathbb{I}(T_2 \le C_2)$ with $C_2 = (C - T_1)\mathbb{I}(T_1 \le C)$ the censoring variable of the second gap time. Define $\tilde{Y} = Y \wedge C$ and let F_1, F, G and H denote the distribution functions of T_1, Y, C and $\tilde{Y}$, respectively. Since T_1 and C are independent, the Kaplan-Meier product-limit estimator based on the pairs $\left(\tilde{T}_{1i}, \Delta_{1i}\right)$'s , consistently estimates the distribution F_1. Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on $\left(\tilde{T}_{1i} + \tilde{T}_{2i}, \Delta_{2i}\right)$'s. Because T_2 and C_2 will be in general dependent, the estimation of the marginal distribution for the second gap time is not a simple issue. The same applies to the bivariate distribution function $F_{12}(x, y) = \mathbb{P}(T_1 \le x, T_2 \le y)$.This issue have received much attention recently. Among others it was investigated by Wang and Wells [9], Lin et al. [2], Wang and Chang [8], Peña et al. [4], van der Laan et al. [6] or van Keilegom [7].

In this work we present a simple estimator for the bivariate distribution function of the gap times. This estimator is somehow related (although not equal) to that proposed in [2]. The idea behind the estimator is using the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. Some related problems as estimation of the marginal distribution of the second gap time will be discussed. Simulation studies were conducted to

assess the properties of the proposed estimator. We applied the proposed methods to a study of colon cancer data [3].

2. The estimator

Introduce

$$\widehat{F}_{12}(x,y)=\sum_{i=1}^{n}W_i\mathbb{I}\left(\tilde{T}_{1i}\le x,\tilde{T}_{2i}\le y\right) \tag{1}$$

where $W_i=\frac{\Delta_{2i}}{n-R_i+1}\prod_{j=1}^{i-1}\left[1-\frac{\Delta_{2j}}{n-R_j+1}\right]$, is the Kaplan-Meier weight attached to $\tilde{Y}_i$ when estimating the marginal distribution of Y from $\left(\tilde{Y}_i,\Delta_{2i}\right)$'s, and for which the ranks of the censored $\tilde{Y}_i$'s, R_i, are higher than those for uncensored values in the case of ties.

This estimator is consistent (shown below) whenever $x+y$ is smaller than the upper bound of the support of the censoring time. From (1) we can obtain an estimator for the marginal distribution of the second gap time, $F_2(y)=\mathbb{P}(T_2\le y)$, namely

$$\widehat{F}_2(y)=\widehat{F}_{12}(\infty,y)=\sum_{i=1}^{n}W_i\mathbb{I}\left(\tilde{T}_{2i}\le y\right) \tag{2}$$

Note that estimator (2) is not the Kaplan-Meier estimator based on $\left(\tilde{T}_{2i},\Delta_{2i}\right)$'s. This is because the weights W_i are based on the $\tilde{Y}_i$-ranks rather than on the $\tilde{T}_{2i}$-ranks. Indeed, since T_2 and C_2 are expected to be dependent, the ordinary Kaplan-Meier estimator of F_2 will be in general inconsistent.

Let τ_F be the upper bound of the support of F, an similarly define τ_G and τ_H. From the independence assumption, we have $\tau_H=\tau_F\wedge\tau_G$. Let *A* be the (possibly empty) set of atoms of $\tilde{Y}$. We have the following result.

Theorem 1 If F and G have no jumps in common, we have with probability 1 and in the mean

$$\lim_{n\to\infty}\widehat{F}_{12}(x,y)=\mathbb{P}\left(T_1\le x,T_2\le y,T_1+T_2\le\tau_H\right)+\mathbb{I}\left(\tau_H\in A\right)\mathbb{P}\left(T_1\le x,T_2\le y,T_1+T_2=\tau_H\right),$$

and

$$\lim_{n\to\infty}\widehat{F}_2(y)=\mathbb{P}\left(T_2\le y,T_1+T_2\le\tau_H\right)+\mathbb{I}\left(\tau_H\in A\right)\mathbb{P}\left(T_2\le y,T_1+T_2=\tau_H\right),$$

Proof. Use Stute [5].

Note that if any of the following conditions hold:

(a) $\tau_H\in A$, or

(b) $\mathbb{P}\left(T_1\le x,T_2\le y,T_1+T_2=\tau_H\right)=0$

we have from Theorem 1 that $\lim_{n\to\infty}\widehat{F}_{12}(x,y)=\mathbb{P}\left(T_1\le x,T_2\le y,T_1+T_2\le\tau_H\right)$. Condition (b) holds in particular if *Y* is continuous. Then three different situations are possible:

(A) If $\tau_F<\tau_G$ (or if $\tau_F=\tau_G=\infty$), than we get consistency for (1) for any (x,y).

(B) If $\tau_G<\tau_F$, then $\tau_H<\tau_F$ and consistency is only ensured for $x+y<\tau_H$ (or for $x+y\le\tau_H$ provided that (a) or (b) above hold).

(C) If $\tau_F=\tau_G<\infty$ then consistency follows if (a) or (b) is fulfilled.

The estimator (1) is somehow related (but not equal) to that proposed in Lin et al. [2]. In fact, in the next Section we report a simulation study to compare both estimators. The estimator proposed in Lin's paper is expressed as

$$\widetilde{F}_{12}(x,y) = \widetilde{H}(x,0) - \widetilde{H}(x,y)$$

where

$$\widetilde{H}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}\left(\tilde{T}_{1i} \le x, \tilde{T}_{2i} > y\right)\Delta_{2i}}{1-\hat{G}\left(\left(\tilde{T}_{1i}+y\right)^{-}\right)}$$

and where $\widehat{G}$ stands for the Kaplan-Meier estimator based on the $\left(\tilde{Y}_i, 1-\Delta_{2i}\right)$'s. The estimator (1) proposed here can also be written as $\widehat{F}_{12}(x,y) = \widehat{H}(x,0) - \widehat{H}(x,y)$ where

$$\widehat{H}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}\left(\tilde{T}_{1i} \le x, \tilde{T}_{2i} > y\right)\Delta_{2i}}{1-\hat{G}\left(\left(\tilde{T}_{1i}+\tilde{T}_{2i}\right)^{-}\right)}.$$

3. Simulation study

In this section, we compare by simulations the new estimator, $\widehat{F}_{12}(x,y)$, for the bivariate distribution function to that proposed in Lin et al. [2], $\widetilde{F}_{12}(x,y)$. The simulated scenario is the same as that described in Lin's paper (see their Section 3). In this scenario, the gap times were generated from Gumbel's bivariate distribution function

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

where the marginal distribution functions F_1 and F_2 are exponential with rate parameter 1. The parameter θ was set to 0 for simulating independent gap times, and also to 1, corresponding to 0.25 correlation between the two gap times. The censoring times were generated from a uniform according to model $U(0,3)$. For each simulation 10000 samples were generated with sample size of 100. Other sample sizes (n=50) and level of censoring (according to model $U(0,4)$) were considered although the results are not shown here. For each setting we computed the mean and standard deviations for the bivariate estimator $\widehat{F}_{12}(x,y)$ at pairs of time points (x,y), where x and y takes values 0.2231, 0.5108, 0.9163 and 1.6094, corresponding to marginal survival probabilities of 0.8, 0.6, 0.4 and 0.2. The true values of $F_{12}(x,y)$ are reported in Table 1.

Table 2 reports the mean estimate along with the corresponding standard deviation. As it can be seen, the bias of the bivariate distribution estimator achieved reasonable levels. The variance increases at the right tail of the bivariate distribution, where the censoring effects are stronger. We also computed the bias and the deviation of the estimator proposed in Lin et al. [2]. The bias turned out to be of the same order as that of (1). Table 3 reports the efficiency of $\widetilde{F}_{12}(x,y)$ relative to $\widehat{F}_{12}(x,y)$. This efficiency was measured through the squared quotient of standard deviations. We see that $\widehat{F}_{12}(x,y)$ was always more efficient than $\widetilde{F}_{12}(x,y)$ except for a few cases corresponding to small *x* and large *y*.

	$\theta = 0$					$\theta = 1$			
$x \setminus y$	0.2231	0.5108	0.9163	1.6094		0.2231	0.5108	0.9163	1.6094
0.2231	0.0400	0.0800	0.1200	0.1600		0.0656	0.1184	0.1584	0.1856
0.5108	0.0800	0.1600	0.2400	0.3200		0.1184	0.2176	0.2976	0.3584
0.9163	0.1200	0.2400	0.3600	0.4800		0.1584	0.2976	0.4176	0.5184
1.6094	0.1600	0.3200	0.4800	0.6400		0.1856	0.3584	0.5184	0.6656

Table 1. True values of the bivariate distribution of the gap times under the simulated model

	$\theta = 0$					$\theta = 1$			
$x \setminus y$	0.2231	0.5108	0.9163	1.6094		0.2231	0.5108	0.9163	1.6094
0.2231	0.0398 (.0204)	0.0801 (.0287)	0.1209 (.0355)	0.1604 (.0435)		0.0652 (.0254)	0.1176 (.0340)	0.1580 (.0400)	0.1853 (.0444)
0.5108	0.0801 (.0287)	0.1600 (.0399)	0.2404 (.0490)	0.3203 (.0584)		0.1184 (.0344)	0.2170 (.0454)	0.2978 (.0520)	0.3577 (.0596)
0.9163	0.1203 (.0361)	0.2396 (.0495)	0.3598 (.0580)	0.4798 (.0679)		0.1587 (.0397)	0.2985 (.0518)	0.4179 (.0598)	0.5176 (.0668)
1.6094	0.1599 (.0434)	0.3197 (.0579)	0.4805 (.0680)	0.6345 (.0786)		0.1858 (.0439)	0.3587 (.0582)	0.5176 (.0669)	0.6581 (.0773)

Table 2. Mean value and standard deviation of $\widehat{F}_{12}(x, y)$ along 10,000 simulated samples.

	$\theta = 0$					$\theta = 1$			
$x \setminus y$	0.2231	0.5108	0.9163	1.6094		0.2231	0.5108	0.9163	1.6094
0.2231	0.7867	0.8627	0.9010	0.9819		0.8850	0.9383	1.0101	1.1229
0.5108	0.7703	0.8076	0.8452	0.9380		0.8415	0.8983	0.9590	1.0487
0.9163	0.7318	0.7758	0.8086	0.8351		0.7749	0.8317	0.8731	0.8778
1.6094	0.6705	0.6783	0.6355	0.6876		0.6966	0.6854	0.6608	0.6312

Table 3. Efficiency of $\widetilde{F}_{12}(x, y)$ relative to $\widehat{F}_{12}(x, y)$ along the 10,000 simulated samples.

4. A real data example

Due to large number of peoples affected by cancer of colon, there is much demand for information on this disease. In an large percentage of the patients, the diagnosis is made at a sufficiently early stage when all apparent disease tissue can be surgically removed. Unfortunately, some of these patients have residual cancer, which leads to recurrence of disease and death (in some cases). Cancer patients who have experienced a recurrence are known to be at a substantially higher risk of mortality. For the colon cancer data, we may consider the recurrence as an associated state of risk, and use the three-state multi-state model with states "alive and disease-free", "alive with recurrence" and "death". Under gap times framework, T_1 is the time from randomization to cancer recurrence and T_2 is the time from cancer recurrence to death.

In a large clinical trial on Duke's stage III patients (see Moertel et al. [3] for more details), subjects underwent a curative surgery for colo-rectal cancer. From the total of 929 patients, 467 developed recurrence and among these 413 died. In colo-rectal cancer, as in other cancer diseases, is important to make long-terms predictions and to identify possible times of diagnosis (threshold values). In such cases, it is very important to obtain good estimates for the survival probabilities (for the bivariate distribution and for the marginal distributions of the gap times).

Table 4 presents the estimates for the joint distribution function using estimator (1) for several values of (x, y) (the x values are the percentiles 5%, 25% and 50% of the first gap time). Figure 1 illustrate the differences between the Kaplan-Meier estimator for the marginal distribution of the second gap time (based on the $\left(\tilde{T}_{2i}, \Delta_{2i}\right)$'s) and estimator (2). The range of time has been restricted to 550 days to emphasize the differences between the two estimators. Differences between the two curves can be explained by the (possible) failure of the

independence assumption, necessary to obtain consistency for the Kaplan-Meier estimator. Estimates for the two marginal distribution functions (using the Kaplan-Meier product-limit for the first gap time) can be used to compare the survival in two or more groups/treatments (results not shown).

$x \setminus y$	20	100	200	500	1200
56	0.0067	0.0533	0.1333	0.2200	0.2200
346	0.0067	0.1733	0.4934	0.8536	0.8536
1531	0.0133	0.2133	0.5401	0.9136	0.9136

Table 4: Estimates of the joint distribution function, $\widehat{F}_{12}(x, y)$, for the colon cancer study.

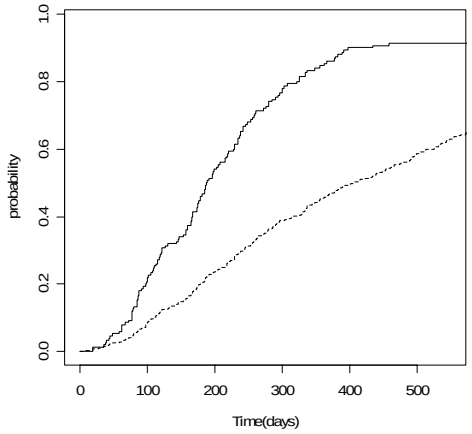

Figure 1: Estimates of the marginal distribution function of the second gap time using $\widehat{F}_2(y)$ (solid line) and using Kaplan-Meier estimator (dashed line). Colon cancer study.

5. Conclusions

In this paper we propose a nonparametric estimator of the bivariate distribution function for censored gap times. In contrast to other existing methods, the introduced estimate is a proper distribution function, in the sense that it attaches positive mass to each observation. We use this estimator to introduce also an estimator for the marginal distribution of the second gap time. Simulations showed that the new estimator is virtually unbiased and that it may achieve efficiency levels clearly above those corresponding to previous proposals. For illustration purposes we used a real dataset from a clinical trail for colon cancer.

Acknowledgement. The authors acknowledge finantial support by Spanish Ministry of Education & Science grants MTM2005-01274 and MTM2005-00818 (European FEDER support included), and by the Xunta de Galicia grants PGIDIT06PXIC208043PN and PGIDIT06PXIC300117PN.

6. References

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

[2] Lin, D.Y., Sun, W. and Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* 86, 59-70.

[3] Moertel, C.G., Fleming T.R., McDonald J.S. et al. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New Eng. J. Med.* 352-358.

[4] Peña, E.A., Strawderman, R.L. and Hollander, M. (2001). Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association* 96, 1299-1315.

[5] Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of the Multivariate Analysis* 45, 89-103.

[6] van der Laan, M.J., Hubbard, A.E. and Robins, J.M. (2002). Locally efficient estimation of a multivariate survival function in longitudinal studies. *Journal of the American Statistical Association* 97, 494-507.

[7] van Keilegom, I. (2004). A note on the nonparametric estimation of the bivariate distribution under dependent censoring. *Journal of Nonparametric Statistics*, 16, 659-670.

[8] Wang, M.C. and Chang, S.H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association* 94, 146-153.

[9] Wang, W. and Wells, M.T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika* 85, 561-572.