Estimation of multivariate distributions for recurrent event data

Luís Meira-Machado[1], Beatriz Sampaio[1]

[1] Centre of Mathematics & Department of Mathematics and Applications, University of Minho, Campus de Azurem, 4800-058 Guimarães, Portugal.

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

Abstract: In many longitudinal studies information is collected on the times of different kinds of events. Some of these studies involve repeated events, where a subject or sample unit may experience a well-defined event several times along his history. Such events are called recurrent events. In this work we consider the estimation of the marginal and joint distribution functions of two gap times under univariate random right censoring. We also consider the estimation of the bivariate survival function.

Keywords: Censoring; Kaplan-Meier; Nonparametric estimation; Recurrent events; Survival Analysis.

1 Introduction

In many longitudinal studies, subjects can experience recurrent events. This type of data has been frequently observed in medical research, engineering, economy and sociology. In medical research, the recurrent events could be multiple occurrences of hospitalization from a group of patients, multiple recurrence episodes in cancer studies, repeated heart attacks or multiple relapses from remission for leukemia patients. In this work we consider the estimation of the marginal and joint distribution / survival functions of the gap times under univariate random right censoring. These issues have received much attention recently. Among others they were investigated by Lin, Sun and Ying (1999), de Uña-Álvarez and Meira-Machado (2008), de Uña-Álvarez and Amorim (2011) or Moreira, Araújo and Meira-Machado (2017).

2 Nonparametric estimators

In the context of recurrent event data, each individual may go through a well-defined event several times along his history. Assume that each study subject can potentially experience K consecutive events at times $T_1 < T_2 <$

$... < T_K$, which are measured from the start of the follow-up. In this work we are primarily interested in the gap times $Y_1 := T_1$, $Y_2 := T_2 - T_1$, ..., $Y_k := T_k - T_{k-1}$, $k = 2, ..., K$. For simplicity we assume $K = 2$.
Then, (Y_1, Y_2) is a vector of gap times of successive events, which we assume to be observed subjected to (univariate) random right-censoring. Let C be the right-censoring variable, assumed to be independent of (Y_1, Y_2). Because of this, the observed data consists of $(\widetilde{Y}_{1i}, \widetilde{Y}_{2i}, \Delta_{1i}, \Delta_{2i})$, $1 \leq i \leq n$, which are n independent replications of $(\widetilde{Y}_1, \widetilde{Y}_2, \Delta_1, \Delta_2)$, where $\widetilde{Y}_1 = Y_1 \wedge C$, $\Delta_1 = I(Y_1 \leq C)$, $\widetilde{Y}_2 = Y_2 \wedge C_2$, $\Delta_2 = I(Y_2 \leq C_2)$ with $C_2 = (C - Y_1)I(Y_1 \leq C)$ the censoring variable of the second gap time. Here and thereafter, $a \wedge b = min(a, b)$ and $I(\cdot)$ is the indicator function.
Let F_k, $k = 1, 2$ denote the distribution function of the k-th event time T_k. Since T_k and C are independent, the Kaplan-Meier product-limit estimator (Kaplan and Meier, 1958) based on the pairs $(\widetilde{T}_{ki}, \Delta_{ki})$'s, consistently estimates the distribution of the time to the k-th event. Because Y_2 and C_2 will be in general dependent, the estimation of the marginal distribution of the second gap time is not a simple issue. The same applies to the joint distribution function $F_{12}(t_1, t_2) = P(Y_1 \leq t_1, Y_2 \leq t_2)$ and the joint survival function $S_{12}(t_1, t_2) = P(Y_1 > t_1, Y_2 > t_2)$. Some estimators for these quantities will be presented below.
Below we present several different approaches for estimating the bivariate distribution function of (Y_1, Y_2). An estimator based on Inverse Probability of Censoring Weights was first introduced by Lin, Sun and Ying (1999):

$$\widehat{F}_{12}^{\texttt{IPCW}}(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(\widetilde{Y}_{1i} \leq t_1)\Delta_{1i}}{\widehat{G}_1(\widetilde{Y}_{1i})} - \frac{1}{n}\sum_{i=1}^{n} \frac{I(\widetilde{Y}_{1i} \leq t_1, \widetilde{Y}_{2i} > t_2)}{\widehat{G}(\widetilde{Y}_{1i} + t_2)}.$$

where $\widetilde{G}_1$ and $\widetilde{G}$ stand for the Kaplan-Meier estimator (of the censoring distribution) based on the $(\widetilde{Y}_{1i}, 1 - \Delta_{1i})$'s and $(\widetilde{T}_{2i}, 1 - \Delta_{2i})$'s, respectively.
A simple estimator based on the Kaplan-Meier weights was later introduced by de Uña-Álvarez and Meira-Machado (2008). The idea behind their estimator is to weight the data by the Kaplan-Meier weights (W_i) pertaining to the distribution of the total time (in this case, T_2) of the process:

$$\widehat{F}_{12}^{\texttt{KMW}}(t_1, t_2) = \sum_{i=1}^{n} W_i I(\widetilde{Y}_{1i} \leq t_1, \widetilde{Y}_{2i} \leq t_2).$$

A related estimator based on presmoothing ($\widehat{F}_{12}^{\texttt{PKMW}}$) was later proposed by de Uña-Álvarez and Amorim (2011). Successful applications of presmoothed estimators include nonparametric curve estimation, regression analysis and estimation of the transition probabilities (Moreira et al. 2013).
Given that $P(Y_1 \leq t_1, Y_2 \leq t_2) = P(Y_2 \leq t_2 \mid Y_1 \leq t_1)P(Y_1 \leq t_1)$ we also consider the landmark estimator (LDM) for which to estimate $P(Y_2 \leq t_2 \mid$

$Y_1 \leq t_1$) the analysis is restricted to the individuals with an observed first event time less or equal than t_1. This is known as the landmark approach (van Houwelingen et al. 2007). The corresponding estimator (LDM) is given by

$$\widehat{F}_{12}^{\texttt{LDM}}(t_1, t_2) = \sum_{i=1}^{n} W_i^{(t_1]} I(\widetilde{Y}_{2i} \leq t_2) \times \widetilde{F}_1^{KM}(t_1)$$

where F_1^{KM} is the Kaplan-Meier estimator of the distribution of the first time and $W_i^{(t_1]}$ denote the Kaplan-Meier weights of the distribution of T_2 computed from the given sub sample $\left\{i : \widetilde{Y}_1 \leq t_1\right\}$.

In this work we also introduce new estimators which are constructed using the cumulative hazard of the total time given a first time but where each observation has been weighted using the information of the first duration. The proposed estimator (WCH - weighted cumulative hazard) is given by $\widehat{F}_{12}^{\texttt{WCH}}(t_1, t_2) = \widehat{P}(Y_1 \leq t_1)(1 - \widehat{P}(Y_2 > t_2 \mid Y_1 \leq t_1))$ where $\widehat{P}(Y_1 \leq t_1)$ is estimated by the Kaplan-Meier estimator of the first event time and $\widehat{P}(Y_2 > t_2 \mid Y_1 \leq t_1) = \prod_{v \leq t_2}(1 - \hat{\Lambda}_{Y_2|Y_1 \leq t_1}(dv))$,
where

$$\hat{\Lambda}_{Y_2|Y_1 \leq t_1}(dv) = \frac{\sum_{i=1}^{n} I(\widetilde{Y}_{1i} \leq t_1, \widetilde{Y}_{2i} = v, \Delta_{2i} = 1)/\widehat{G}(\widehat{Y}_{1i} + v)}{\sum_{i=1}^{n} I(\widetilde{Y}_{1i} \leq t_1, \widetilde{Y}_{2i} \geq v, \Delta_{1i} = 1)/\widehat{G}(\widehat{Y}_{1i} + v)}.$$

Finally we compare the aforementioned methods with the estimator of the bivariate distribution which is obtained using Nearest Neighbor Estimation (NNE).

Now, we consider the estimation of the bivariate survival function $S(t_1, t_2) = P(Y_1 > t_1, Y_2 > t_2)$. For this quantity, the estimator constructed using the Kaplan-Meier weights was built assuming the following equality $S(t_1, t_2) = 1 - P(Y_1 \leq t_1) - P(Y_1 > t_1, Y_2 \leq t_2)$ where the first probability on the right hand side is estimated using the Kaplan-Meier estimator of the first event and the second probability is estimated using Kaplan-Meier weights pertaining to the distribution of the total time (i.e., T_2) in a similar way as introduced above. The weighted cumulative hazard estimator of the bivariate survival function is given by $\widehat{S}_{12}^{\texttt{WCH}}(t_1, t_2) = \widehat{P}(Y_2 > t_2 \mid Y_1 > t_1)(1 - \widehat{P}(Y_1 \leq t_1))$ where $\widehat{P}(Y_2 > t_2 \mid Y_1 > t_1)$ is obtained using the same ideas given above. This is the Wang and Wells (1998) estimator.

Finally, landmark-based estimators can be introduced to estimate the bivariate survival function. Given that $P(Y_1 > t_1, Y_2 > t_2) = 1 - P(Y_2 \leq t_2 \mid Y_1 > t_1)(1 - P(Y_1 \leq t_1))$ the idea is to estimate $P(Y_2 \leq t_2 \mid Y_1 > t_1)$ by restricting the analysis to the individuals with an observed first event time greater or equal than t_1. The corresponding estimator (LDM) is given by $\widehat{S}_{12}^{\texttt{LDM}}(t_1, t_2) = 1 - \sum_{i=1}^{n} W_i^{[t_1)} I(\widetilde{Y}_{2i} \leq t_2) \times (1 - \widetilde{F}_1^{KM}(t_1))$ where $W_i^{[t_1)}$ denote the Kaplan-Meier weights of the distribution of T_2 computed from the sub sample $\left\{i : \widetilde{Y}_1 > t_1\right\}$.

3 Example of Application

Our methodology is motivated by the re-analysis of the German breast cancer data. In this study, patients were followed from the date of breast cancer diagnosis until censoring or dying from breast cancer. From the total of 686 women, 299 developed a recurrence and 171 died. These data can be viewed as arising from a model with two consecutive events: 'Alive with Recurrence' and 'Dead'. In this section, we present plots for the proposed methods to estimate the bivariate distribution function and bivariate survival function of the two gap times, Y_1 = "Time to recurrence" and Y_2 = "Time from recurrence to death".
Figure 1 reports estimated probabilities for a fixed value of $t_1 = 365$ (days), along time. Plot shown in the left hand side (bivariate d.f) show that all proposed methods behave quite similar something that is not true with regard to the estimation of the bivariate survival function (right hand side).

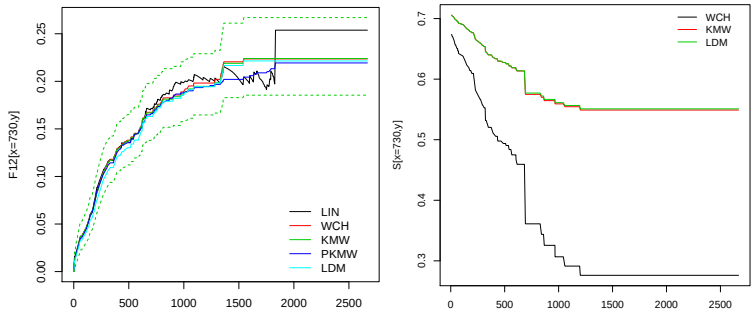

FIGURE 1. Estimates of the bivariate d.f. and bivariate s.f. using the proposed methods. Breast cancer data.

4 Simulation Studies

In this section, we investigate the performance of the proposed estimators through simulations. To simulate the data we consider the bivariate exponential distribution with marginal exponentials with rate parameter 1. This corresponds to the so-called FarlieGumbelMorgenstern copula, where the single parameter controlling for the amount of dependence between the gap times (Moreira and Meira-Machado, 2012; de Uña-Álvarez and Meira-Machado, 2015; Araujo et al., 2015). An independent uniform censoring time C was generated, according to models $Uniform(0, 4)$ and $Uniform(0, 3)$. For each simulated setting we derive the analytic expression of $F_{12}(t_1, t_2)$ and $S_{12}(t_1, t_2)$ for several (t_1, t_2) pairs, corresponding to combinations of the percentiles 20%, 40%, 60%, and 80% of the marginal distributions of the gap times (i.e., 0.2231, 0.5108, 0.9163, 1.6094). Sample sizes $n = 100$, $n = 250$, and $n = 500$ were considered.

Results reveal that the all proposed methods for estimating the bivariate distribution function perform quite well, though the performance of all methods is poorer at the right tail (i.e., larger values of t_1 and t_2) where the censoring effects are stronger. At these points the standard deviation (SD) is in most cases larger. The SD decreases with an increase in the sample size and with a decrease of the censoring percentage. All methods proposed in this work obtain in all settings a negligible bias.
Attained results for the bivariate survival function reveal that the weighted cumulative hazard estimator (WCH) is the recommended approach. This is illustrated in Figure 2 in which we show the boxplots of the estimates for the bivariate distribution function.

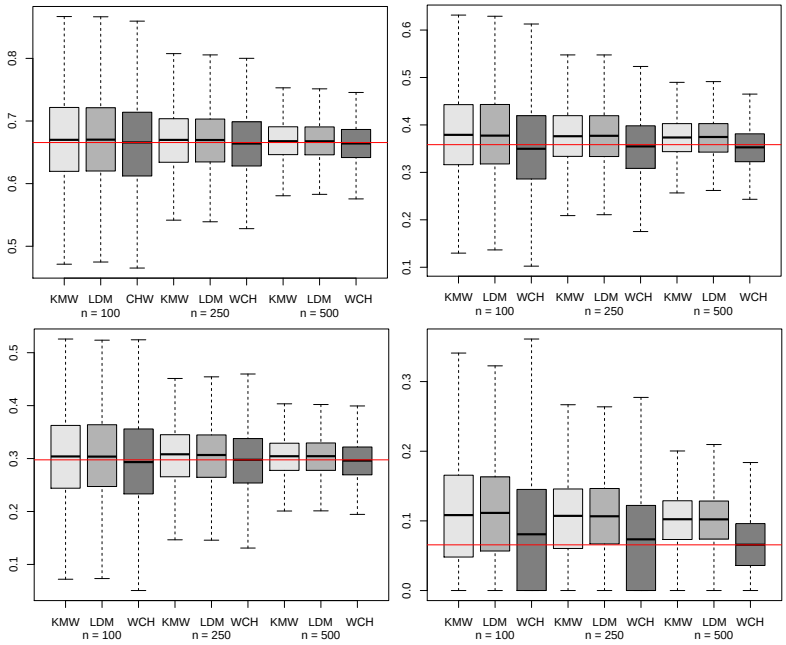

FIGURE 2. Boxplot with estimated probabilities $S_{12}(t_1, t_2)$. On the top results for the pair (0.2231, 0.2231) (left) and (0.2231,0.9163) (right); on the bottom results for the pair (0.9163, 0.5108) (left) and (1.6094,1.6094) (right).

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

References

Araújo A, Meira-Machado L, Roca-Pardiñas J (2015). TPmsm: Estimation of the transition probabilities in 3-state models. *Journal of Statistical Software*, **62(4)**.

de Uña-Álvarez 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.

de Uña-Álvarez 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-Álvarez J, Meira-Machado L. (2015). Nonparametric Estimation of Transition Probabilities in the Non-Markov Illness-Death Model: A Comparative Study. *Biometrics*, **71**, 364 – 375.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53(282)**, 457 – 481.

Lin, D., Sun, W. and Ying, Z. (1999). Nonparametric estimation of the time distributions for serial events with censored data. *Biometrika*, **86(1)**, 59 – 70.

Meira-Machado L, de Uña-Álvarez J, Somnath D. (2015). Conditional Transition Probabilities in a non-Markov Illness-death Model. *Computational Statistics*, **30(2)**, 377 – 397.

Moreira, A. and Meira-Machado, L. (2012) survivalBIV: Estimation of the Bivariate Distribution Function for Sequentially Ordered Events Under Univariate Censoring. *Journal of Statistical Software*, **46(13)**, 1 – 16.

Moreira, A., de Ua-lvarez, J. and Meira-Machado, L. (2013) Presmoothing the Aalen-Johansen estimator of transition probabilities. *Electronic Journal of Statistics*, **7**, 1491 – 1516.

Moreira, A., Araújo, A. and Meira-Machado, L. (2017) Estimation of the bivariate distribution function for censored gap times. *Communications in Statistics - Simulation and Computation*, **46(1)**, 275 – 300.

van Houwelingen, H.C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, **34**, 70 – 85.

Wang, M.C. and Wells, M.T. (1998). Nonparametric Estimation of successive duration times under dependent censoring. *Biometrika*, **85**, 561 – 572.