



# Nonparametric estimation of the distribution of gap times for recurrent events

Gustavo Soutinho<sup>1</sup> · Luís Meira-Machado<sup>2</sup>

Accepted: 18 April 2022 / Published online: 11 May 2022  
© Springer-Verlag GmbH Germany, part of Springer Nature 2022

## 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 throughout their history. Such events are called recurrent events. In this paper, we introduce nonparametric methods for estimating the marginal and joint distribution functions for recurrent event data. New estimators are introduced and their extensions to several gap times are also given. Nonparametric inference conditional on current or past covariate measures is also considered. We study by simulation the behavior of the proposed estimators in finite samples, considering two or three gap times. Our proposed methods are applied to the study of (multiple) recurrence times in patients with bladder tumors. Software in the form of an R package, called `survivalREC`, has been developed, implementing all methods.

**Keywords** Censoring · Gap times · Kaplan–Meier · Multiple events · Recurrent events

## 1 Introduction

In many longitudinal studies, subjects can experience recurrent events (Cook and Lawless 2007). This type of data has been frequently observed in medical research, engineering, the economy, and sociology. In medical research, recurrent events could be multiple occurrences of hospitalization for a group of patients, multiple recurrence

---

✉ Gustavo Soutinho  
gdsoutinho@gmail.com

Luís Meira-Machado  
lmachado@math.uminho.pt

<sup>1</sup> EPIUnit, University of Porto, Rua das Taipas 135, 4050-600 Porto, Portugal

<sup>2</sup> Centre of Mathematics and Department of Mathematics, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal

episodes in cancer studies, recurrent upper respiratory and ear infections, repeated heart attacks, or multiple relapses from remission for leukemia patients (Byar 1980; Pepe and Cai 1993; Wei et al. 1989). The analysis of such data can be focused on time-between-events (gap times) or time-to-event models. In time-to-event models, the events of concern usually represent different states in the disease process (e.g., alive and disease-free, alive with disease, and dead) and they are modeled through their intensity functions (Andersen et al. 1993; Meira-Machado et al. 2009; Meira-Machado and Sestelo 2019). In these models, the estimation of these quantities is essential for long-term survival prognosis. For instance, in cancer studies, one could consider the time to recurrence and the time to death for recurrent patients as the gap times. Under this setting, many other medical contexts can be found in the literature, such as asthma, HIV/AIDS, heart disease, dementia, Alzheimer's disease, etc. Though the proposed methods can be used in both settings, in this paper, we consider that the events are of the same nature and focus on time-between-events. This line of research has received much attention recently. Among others, they were investigated by Campbell (1981), Burke (1988), Lin et al. (1999), Van Keilegom (2004), Peña et al. (2001), de Uña-Álvarez and Meira-Machado (2008), de Uña-Álvarez and Amorim (2011) and Moreira et al. (2017) whose interest was focused on the estimation of the bivariate distribution of the gap times. In other cases, the interest was more focused on the distribution of the gap times, such as the estimation of the joint gap times, the gap time survival functions, or the conditional survival function of the gap times (Meira-Machado et al. 2016; Meira-Machado and Sestelo 2016). Among others, these issues were investigated by Tsai et al. (1986), Prentice and Cai (1992), Lin and Ying (1993), van der Laan et al. (2002), Wang and Wells (1998), Wang and Chang (1999), Prentice et al. (2004) and Schaubel and Cai (2004). These approaches are focused on a pair of gap times corresponding to two consecutive events, and the extension to cope with a vector of  $k$  gap times may not be obvious. Furthermore, the proposed methods do not account for the influence of covariates. In addition, the implementation of several of the aforementioned methods will be difficult in practice due to the lack of user friendly software. The present paper aims to fill this gap. We consider the nonparametric estimation of the multivariate distribution functions of the gap times under univariate random right censoring conditionally (or not) on current or past covariate measures. New estimators for  $K \geq 2$  gap times are introduced and their performances and limitations are discussed. One set of estimators considers a subsampling approach—which we term landmark (de Uña-Álvarez and Meira-Machado 2015)—where a selection is made of the data consisting of subjects occupying a given state at a particular time. Alternative weighted cumulative hazard estimators are also proposed. The idea is to use an adaptation of the nonparametric estimator presented by Wang and Wells (1998) which is 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 methods can also be used to obtain conditional probabilities such as those provided in the plots shown in section 4 that provide useful interpretation. We also introduce a feasible estimation method for the multivariate distribution function, conditionally on covariate measures. The proposed method follows the ideas of Meira-Machado et al. (2015) in which the authors use kernel

weights and the principle of ‘inverse probability of censoring weighting’ (IPCW) (to estimate these quantities conditionally on a continuous covariate. Finally, a tutorial for analyzing such types of data using an R package in which all methods are implemented.

It is worth mentioning that there are several modelling techniques for analyzing the effect of covariates in recurrent time-to-event data. To that end, extensions of the proportional hazards model, such as the Andersen and Gill model (AG) (Andersen and Gill 1982), the Prentice, Williams, and Peterson (PWP) model (Prentice et al. 1981), and Wei, Lin, and Weissfeld (WLW) (Wei et al. 1989), have been proposed for analyzing recurrent event data. An overview of these methods can be seen in the paper by Amorim and Cai (2015) and they are outside the scope of this paper.

This article is organized as follows. The next section presents the notation and introduces the estimators. The finite sample properties of the estimators are studied by simulation in Sect. 3. In Sect. 4 we give a brief overview of the `survivalREC` R package developed by the authors. We illustrate how these methods can be used for data exploration by applying them to a data set on bladder cancer. Main conclusions and discussion are reported in Sect. 5.

## 2 Estimators

### 2.1 Notation

In the context of recurrent event data, each individual may go through a well-defined event several times in his history. Assume that each study subject can potentially experience  $K$  consecutive events at times  $T_1 < T_2 < \dots < T_K$ , which are measured from the start of the follow-up. We are primarily interested in the gap times  $Y_1 := T_1$ ,  $Y_2 := T_2 - T_1, \dots, Y_k := T_k - T_{k-1}, k = 2, \dots, K$ . Then,  $T_k = Y_1 + \dots + Y_k$  is the time to the  $k$ th event,  $T_K$  is the total time and  $(Y_1, Y_2, \dots, Y_K)$  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  $(T_1, T_2, \dots, T_K)$ . Because of this, the observed data consists of  $(\tilde{Y}_{1i}, \dots, \tilde{Y}_{Ki}, \Delta_{1i}, \dots, \Delta_{Ki}), 1 \leq i \leq n$ , which are  $n$  independent replications of  $(\tilde{Y}_1, \dots, \tilde{Y}_K, \Delta_1, \dots, \Delta_K)$ , where  $\tilde{Y}_1 = Y_1 \wedge C, \Delta_1 = I(Y_1 \leq C), \tilde{Y}_2 = Y_2 \wedge C_2, \Delta_2 = I(Y_2 \leq C_2)$  with  $C_2 = (C - Y_1)I(Y_1 \leq C)$  the censoring variable of the second gap time and  $\tilde{Y}_k = Y_k \wedge C_k, \Delta_k = I(Y_k \leq C_k)$  with  $C_k = (C - Y_{k-1})I(Y_{k-1} \leq C)$ . Obviously,  $\Delta_k = 1$  implies  $\Delta_1 = \dots = \Delta_{k-1} = 1$ . Define also  $\tilde{T}_k = T \wedge C$ . Here and thereafter,  $a \wedge b = \min(a, b)$  and  $I(\cdot)$  is the indicator function.

Let  $F_k$  denote the distribution function of the  $k$ th event time  $T_k$  and  $F_k^\star$  denote the distribution function of the  $k$ th gap time  $Y_k$ . Due to the independence assumption between  $C$  and  $(T_1, \dots, T_K)$ , the marginal distribution of the  $k$ th event time can be consistently estimated by the Kaplan–Meier estimator based on the  $(\tilde{T}_k, \Delta_k)$ ’s. Note that, since the variables  $T_1 < T_2 < \dots < T_K$  are recorded successively and are subject to censoring, we only observe the  $k$ th gap time if all previous failure times are uncensored. In practice this will imply that for  $k > 1, Y_k$  and  $C_k$  will be in general

dependent, which will make difficult the estimation of the marginal distribution of the  $k$ th gap time, as well as the estimation to the joint distribution function  $F_{1\dots k}(y_1, \dots, y_k) = P(Y_1 \leq y_1, \dots, Y_k \leq y_k)$ .

### 2.2 Estimators for the bivariate distribution function

In this section we will present different approaches for estimating the bivariate distribution function of  $(Y_1, Y_2)$ ,  $F_{12}(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2)$ . The generalization to  $K > 2$  gap times will be given in a later section.

#### 2.2.1 Inverse probability of censoring weighted estimators

In the absence of censoring, the bivariate distribution function can be empirically estimated by  $\hat{F}_{12}(y_1, y_2) = \frac{1}{n} \sum_{i=1}^n I(\tilde{Y}_{1i} \leq y_1, \tilde{Y}_{2i} \leq y_2)$ . To handle right censoring, inverse probability of censoring weighting (IPCW) can be used (see, for example, Moreira et al. (2017) for further details).

The idea of IPCW was used by Lin et al. (1999) to introduce an estimator for the bivariate distribution function. Their estimator is based on the relation  $P(Y_1 \leq y_1, Y_2 \leq y_2) = P(Y_1 \leq y_1) - P(Y_1 \leq y_1, Y_2 > y_2)$  where the first quantity in the right-hand side of the equation can be consistently estimated using the Kaplan–Meier estimator Kaplan and Meier (1958) of the distribution function of the first event time  $T_1$  (i.e., based on the pairs  $(\tilde{T}_{1i}, \Delta_{1i})$ 's) which we denote by  $\hat{F}_1$ . The idea to estimate the second term follows from the following relation  $E[I(Y_1 \leq y_1, Y_2 > y_2)] = E[\frac{I(Y_1 \leq y_1, Y_2 > y_2)I(C > T_1 + y_2)}{G(T_1 + y_2)}] = E[\frac{I(\tilde{Y}_1 \leq y_1, \tilde{Y}_2 > y_2)}{G(T_1 + y_2)}]$ . From this, it follows that

$$\hat{F}_{12}^{LIN}(y_1, y_2) = \hat{F}_1(y_1) - \frac{1}{n} \sum_{i=1}^n \frac{I(\tilde{Y}_{1i} \leq y_1, \tilde{Y}_{2i} > y_2)}{\hat{G}(\tilde{Y}_{1i} + y_2)} \tag{1}$$

where  $\hat{G}$  stands for the Kaplan–Meier estimator of the censoring distribution which is computed using the  $(\tilde{T}_{2i}, 1 - \Delta_{2i})$ 's.

Later, de Uña- Álvarez and Meira-Machado (2008) proposes an alternative estimator which is defined in terms of multivariate ‘Kaplan–Meier integrals’ with respect to the marginal distribution of  $T_2$ . The idea behind their estimators is to weight the bivariate data using the Kaplan–Meier estimator of  $T_2$  as shown below.

$$\hat{F}_{12}^{KMW}(y_1, y_1) = \sum_{i=1}^n W_i I(\tilde{Y}_{1i} \leq y_1, \tilde{Y}_{2i} \leq y_2). \tag{2}$$

where  $W_i$  is the Kaplan–Meier weight attached to  $\tilde{T}_{2i}$  when estimating the marginal distribution of  $T_2$  from  $(\tilde{T}_{2i}, \Delta_{2i})$ 's (equal to minus the jump at  $\tilde{T}_{2i}$  of the Kaplan–Meier estimator of survival of the total time; see de Uña- Álvarez and Meira-Machado (2008) for more details).

Estimator (2) (labeled as *KMW*) can also be expressed using IPCW (see de Uña-Álvarez and Meira-Machado 2008) being somehow related (although not equal) to that proposed by Lin et al. (1999). The two estimators labeled by *KMW* and *LIN* deal with right censoring using an appropriate reweighting of the chosen summands, and the differences between them are somewhat subtle. The *KMW* estimator only puts mass on observations that are completely uncensored, whereas Lin’s estimator puts mass on observations that were uncensored till a given time. In practice, this means that *LIN* estimator will show estimated curves with more jump points. The estimates produced via the *KMW* estimator produce a valid bivariate distribution since it does guarantee that the bivariate distribution function is monotone. In contrast, the specific reweighting of the data that is used in Lin’s estimator does not ensure this property. Their estimators do not attach positive mass to each pair of recorded gap times, which may lead to problems of interpretation. A proper estimator for the bivariate distribution function could be obtained by keeping the estimator constant until it starts decreasing again. However, this approach provides a downward-biased estimator. These features can be seen in our application section. The two estimators are consistent whenever  $y_1 + y_2$  is smaller than the upper bound of the support of the censoring time.

### 2.2.2 Estimators based on conditional probabilities

In this section, we propose estimators for the bivariate distribution function that consider the relation  $P(Y_1 \leq y_1, Y_2 \leq y_2) = P(Y_2 \leq y_2 | Y_1 \leq y_1)P(Y_1 \leq y_1)$ , where the second term in the right-hand side of the equation can be estimated using the Kaplan–Meier product-limit estimator of the distribution function of the first event time. Two different estimation methods are proposed below to estimate the first term on the right-hand side of the equation shown above.

A simple estimator for the bivariate distribution function considers that the first term can be estimated using a subsampling approach. This approach, which we also term as landmarking (van Houwelingen 2007), is obtained by considering specific subsamples or portions of the data at hand. In this case, for estimating the conditional probability  $P(Y_2 \leq y_2 | T_1 \leq y_1)$ , the analysis is restricted to those individuals with a first gap time (equivalently, the first event time) less or equal to  $y_1$ . To formalize things, let  $n_1$  be the cardinal of  $\mathcal{S} = \{i : \tilde{Y}_{1i} \leq y_1\}$ . Then,

$$\hat{F}^{LDM}(y_2 | Y_1 \leq y_1) = \sum_{i=1}^{n_1} W_i^{(y_1)} I(\tilde{Y}_{2i} \leq y_2). \tag{3}$$

where  $W_i^{(y_1)}$  are the Kaplan–Meier weights of the distribution of  $T_2$  computed from the subsample  $\mathcal{S}$ .

Any of the estimators proposed above (*LIN*, *KMW* and *LDM*) may reveal some problems in the right tail where uncensored observations are scarce. Below, we propose an estimator that may deal more efficiently with those situations. The proposed estimator is 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. This estimator (WCH—weighted cumulative hazard) follows the ideas by Wang and Wells (1998) in which a product-limit estimator for the second gap time is used:

$$\widehat{F}_{12}^{WCH}(y_1, y_2) = \widehat{F}_2^{\star WCH}(y_2 | Y_1 \leq y_1) \widehat{F}_1(y_1) \tag{4}$$

where  $\widehat{F}_1(y_1)$  is the Kaplan–Meier estimator and  $\widehat{F}_2^{\star WCH}(y_2 | Y_1 \leq y_1) = 1 - \prod_{v \leq y_2} (1 - \widehat{\Lambda}_{Y_2 | Y_1 \leq y_1}(dv))$  for which

$$\widehat{\Lambda}_{Y_2 | Y_1 \leq y_1}(dv) = \frac{\sum_{i=1}^n I(\widetilde{Y}_{1i} \leq y_1, \widetilde{Y}_{2i} = v, \Delta_{2i} = 1) / \widehat{G}(\widehat{Y}_{1i} + v)}{\sum_{i=1}^n I(\widetilde{Y}_{1i} \leq y_1, \widetilde{Y}_{2i} \geq v, \Delta_{1i} = 1) / \widehat{G}(\widehat{Y}_{1i} + v)}.$$

One interesting topic in the analysis of recurrent event data is the estimation of the marginal distribution of the gap times ( $F_j^\star(y) = P(Y_j \leq y)$ ). The problem of estimating these functions has not been discussed explicitly in the literature, giving the impression that the Kaplan–Meier estimator is still the estimator of choice. This is not true for  $j > 1$ . Indeed, since  $T_2$  and  $C_2$  are expected to be dependent, the Kaplan–Meier estimator of  $F_2^\star$  based on the  $(\widetilde{Y}_{2i}, \Delta_{2i})$ 's will be in general inconsistent. An estimator for the marginal distribution of the second gap time can be obtained from proposed estimators of the bivariate distribution function since  $F_2^\star(y) = P(Y_2 \leq y) = F_{12}(+\infty, y)$ . To estimate this quantity, we suggest using the WCH, where the corresponding estimator is a simple adaptation of  $\widehat{F}_2^{\star WCH}$ . An alternative approach would be using the KMW method,  $\widehat{F}_2^{\star KMW}(y) = \sum_{i=1}^n W_i I(\widetilde{Y}_{2i} \leq y)$  where  $W_i$  are the Kaplan–Meier weights attached to  $\widetilde{T}_2$  when estimating the marginal distribution of  $T_2$ .

### 2.3 Extension to the general case of K gap times

In this section we extend the results (estimators) proposed in Sect. 2.2 to the case of  $K$  gap times.

Let  $(Y_1, Y_2, \dots, Y_K)$  denote a vector of  $K$  ordered gap times and let  $F_{1\dots K}$  denote the joint distribution function of  $(Y_1, Y_2, \dots, Y_K)$ . The estimator proposed by de Uña-Álvarez and Meira-Machado (2008) can easily be extended to provide a valid estimator for the joint distribution function of  $(Y_1, Y_2, \dots, Y_K)$ . Their estimator is given by

$$\widehat{F}_{1\dots K}^{KMW}(y_1, \dots, y_K) = \sum_{i=1}^n W_i I(\widetilde{Y}_{1i} \leq y_1, \dots, \widetilde{Y}_{Ki} \leq y_K) \tag{5}$$

where  $W_i$  is the Kaplan–Meier weight attached to  $\widetilde{T}_{Ki}$  when estimating the marginal distribution of  $T_K$  from  $(\widetilde{T}_{Ki}, \Delta_{Ki})$ 's.

Lin's estimator Lin et al. (1999) can be easily extended to the general case of  $K$  gap times. Since  $P(Y_1 \leq y_1, \dots, Y_K \leq y_K) = P(Y_1 \leq y_1, \dots, Y_{K-1} \leq y_{K-1}) - P(Y_1 \leq y_1, \dots, Y_{K-1} \leq y_{K-1}, Y_K > y_K)$ , an obvious estimator for the second term is given by

$$\frac{1}{n} \sum_{i=1}^n \frac{I(\tilde{Y}_{1i} \leq y_1, \dots, \tilde{Y}_{(K-1)i} \leq y_{K-1}, \tilde{Y}_{Ki} > y_K)}{\hat{G}_K(\tilde{T}_{(K-1)i} + y_K)} \tag{6}$$

where  $\hat{G}_K$  stands for the Kaplan–Meier estimator of the censoring distribution based on the  $(\tilde{T}_{Ki}, 1 - \Delta_{Ki})$ 's. The first term can be estimated recursively using the same approach.

The extension of the landmark estimator (LDM) to  $K$  gap times is a consequence of Bayes' theorem,

$$\begin{aligned} \hat{F}_{1\dots K}^{\text{LDM}}(y_1, \dots, y_K) &= \hat{F}_K(y_K \mid Y_1 \leq y_1, \dots, Y_{K-1} \leq y_{K-1}) \\ &\quad \times \hat{F}_{K-1}(y_{K-1} \mid Y_1 \leq y_1, \dots, Y_{K-2} \leq y_{K-2}) \\ &\quad \times \dots \times \hat{F}_2(y_2 \mid Y_1 \leq y_1) \hat{F}_1(y_1) \end{aligned}$$

where  $\hat{F}_j(y \mid Y_1 \leq y_1, \dots, Y_{j-1} \leq y_{j-1})$  is estimated using the Kaplan–Meier estimator based on  $(\tilde{T}_j, \Delta_j)$ 's restricted to those individuals in  $S = \{i : \tilde{Y}_1 \leq y_1, \dots, \tilde{Y}_{j-1} \leq y_{j-1}\}$ .

The extension of the weighted cumulative hazard estimator (WCH) to  $K$  gap times follows from the following relation,

$$\begin{aligned} P(Y_1 \leq y_1, \dots, Y_K \leq y_K) &= (1 - P(Y_K > y_K \mid Y_1 \leq y_1, \dots, Y_{K-1} \leq y_{K-1})) \\ &\quad \times P(Y_1 \leq y_1, \dots, Y_{K-1} \leq y_{K-1}), \end{aligned}$$

where the first term in the right-hand side of the equation is estimated using  $\hat{P}(Y_K > y_K \mid Y_1 \leq y_1, \dots, Y_{K-1} \leq y_{K-1}) = 1 - \prod_{v \leq y_K} (1 - \hat{\Lambda}_{Y_K \mid Y_1 \leq y_1, \dots, Y_{K-1} \leq y_{K-1}}(dv))$  and the second term in the right-hand side of the equation is estimated recursively using the ideas explained in Sect. 2.2.

### 2.4 Estimators conditionally on current or past covariate measures

In this section, we will explain how we may introduce nonparametric estimators for the conditional distribution function,  $F_{12}(y_1, y_2 \mid X)$ . In particular, we are interested in estimating these functions for any time  $y_1$  and  $y_2$ , but conditional on a given continuous covariate  $X$  that could either be a baseline covariate or a current covariate that is observed for an individual during the follow-up. Discrete covariates can also be included by splitting the sample for each level of the covariate and repeating the procedures described in the previous sections for each subsample.

To account for covariate effects, one standard method is to consider estimators based on Cox's model (Cox 1972), with the corresponding baseline hazard function estimated by the Breslow's method (Breslow 1972). Flexible nonparametric effects of the covariates on the bivariate distribution function, as those shown in our example of application, can be obtained using an alternative approach which introduces local smoothing by means of kernel weights based on local constant (Nadaraya–Watson) regression (Nadaraya 1965). The proposed method is introduced in a regression setup based on the inverse probability of censoring weighting.

Assume that we have two consecutive gap times  $(Y_1, Y_2)$  and that  $X$  denotes a continuous covariate. Then, the estimation of these functions can be performed via estimating general conditional expectation of type  $E[\varphi(Y_1, Y_2) | X = x]$ , where  $\varphi$  is a general function defined over  $Y_1$  and  $Y_2$ . For instance, in our setting, for the bivariate distribution function,  $\varphi_{u,v}(Y_1, Y_2) = I(Y_1 \leq u, Y_2 \leq v)$  while for the bivariate survival function  $\varphi_{u,v}(Y_1, Y_2) = I(Y_1 > u, Y_2 > v)$ .

In the absence of censoring, to estimate these quantities nonparametrically, we may use kernel smoothing techniques by calculating a local average of the  $\varphi(Y_1, Y_2)$ , that is, as follows:

$$\widehat{E}[\varphi(Y_1, Y_2)|X = x] = \sum_{i=1}^n W(x, X_i, h)\varphi(Y_{1i}, Y_{2i}),$$

where  $W(x, X_i, h)$  is a weight function which corresponds to the Nadaraya–Watson estimator as follows:

$$W(x, X_i, h) = \frac{k((x - X_i)/h)}{\sum_{j=1}^n k((x - X_j)/h)}$$

where  $k$  is a known probability density function (the kernel function) and  $h$  is the bandwidth.

To handle right censoring, inverse probability of censoring weighting can be used. Since,

$$E[I(Y_1 \leq u, Y_2 \leq v)] = E\left[I(\widetilde{Y}_1 \leq u, \widetilde{Y}_2 \leq v)\Delta_2/G_X(T_2)|X\right]$$

where  $G_X$  denotes the conditional survival function of the censoring time  $C$  given the covariate  $X$ , that is,  $G_{X=x}(t) = P(C > t|X = x)$  which may be estimated using Beran’s estimator (Beran 1981),

$$\widehat{G}_x(t) = \prod_{\widetilde{T}_{2i} \leq t, \Delta_{2i}=0} \left[ 1 - \frac{W(x, X_i, h)}{\sum_{j=1}^n I(\widetilde{T}_{2j} \geq \widetilde{T}_{2i})W(x, X_j, h)} \right] \tag{7}$$

where  $W(x, X_i, h)$  are the Nadaraya–Watson weights.

Based on this, we propose the following nonparametric estimator of the conditional bivariate distribution function:

$$\widehat{F}_{12}(y_1, y_2|X = x) = \sum_{i=1}^n W_i(x, X_i, h) \frac{I(\widetilde{Y}_{1i} \leq y_1, \widetilde{Y}_{2i} \leq y_2)\Delta_{2i}}{\widehat{G}_x(\widetilde{T}_{2i})} \tag{8}$$

where  $G_X^0$  stands for an estimator of the conditional distribution  $C | X$ , for example Beran’s estimator (of the censoring survival function) based on the  $(\widetilde{Y}_{1i}, 1 - \Delta_{1i}, X_i)$ ’s.

Though these methods can be extended to a vector of covariates using multivariate kernels and a generalization of Beran’s estimator, some problems arise with the generalization to higher dimensions.



### 3 Simulation studies

In this section, we compare by simulations the estimators introduced in Section 2. We consider two simulated scenarios, the first scenario aims to compare the estimators introduced in Sect. 2.2. More specifically, estimators for the bivariate distribution function ( $F_{12}(y_1, y_2)$ ) labeled as LIN, KMW, WCH and LDM. The second scenario aims to compare the extensions of the same estimators for three gap times.

#### 3.1 Scenario 1: two gap times

Simulating data in longitudinal recurrent survival data requires the joint modeling of two or more random variables (Meira-Machado and Faria 2014). Copulas provide a useful method for deriving joint distributions given the marginal distributions, especially when the variables are non-normal as in the case of time-to-event variables (Soutinho and Meira-Machado 2020).

For the first scenario, we consider the Fairlie–Gumbel–Morgenstern (FGM) system of bivariate distribution with a joint cumulative distribution function of the form:

$$F_{12}(y_1, y_2) = F_1(y_1)F_2(y_2)[1 + \delta\{1 - F_1(y_1)\}\{1 - F_2(y_2)\}]$$

where  $F_1$  and  $F_2$  are the marginal cumulatives which follow a standard exponential and where  $|\delta| \leq 1$  controls the amount of dependency between the two gap times. It has been shown that in this setting, the correlation of the FGM varies between 0 (independent gap times) for  $\delta = 0$  and 0.25 for  $\delta = 1$ . The FGM was also used in the papers by Lin et al. (1999) and Moreira et al. (2017) Moreira and Meira-Machado (2012).

The follow-up time  $C$  was chosen to be uniformly distributed between 0 and an upper limit. In practice, this limit controls the amount of censored observations. An independent uniform censoring time  $C \sim U[0, 4]$  resulted in 25% of censoring on the first gap time  $Y_1$ , and in 46% of censoring on the second gap time  $Y_2$ , for those individuals with  $\delta = 1$ . A second model with  $C \sim U[0, 3]$  increases these censoring levels to 32% and about 60%, respectively. In each simulation, 1000 samples were generated, each with sample sizes of  $n = 100$  and  $n = 250$ .

Table 1 reports the true values of  $F_{12}(y_1, y_2)$  where  $y_1$  and  $y_2$  take 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. At each time point  $(y_1, y_2)$  we computed the mean squared errors for the four estimators. Table 2 reports these values for model  $C \sim U[0, 3]$  with a sample size of  $n = 250$  and correlated gap times ( $\delta = 1$ ). Our results show that all estimators perform quite well, with reasonable low values for the mean square error. All estimators obtained low values for the bias (not shown) and a worst performance in the right tail (i.e., higher values of  $y_1$  and  $y_2$ ) where the censoring effects are stronger. Though not shown here, the results for different sample sizes and different censoring percentages reveal that an increase in the sample size results in smaller variance and therefore a smaller mean square error. Besides, by increasing the censoring percentage, the standard deviation achieved larger values. The standard deviation

**Table 1** True values of the Fairlie–Gumbel–Morgenstern distribution for two dependent gap times with exponential marginal distribution functions with rate parameter 1

$y_2$	$y_1$			
	0.2231	0.5108	0.9163	1.6094
0.2231	0.0656	0.1184	0.1584	0.1856
0.5108	0.1184	0.2176	0.2976	0.3584
0.9163	0.1584	0.2976	0.4176	0.5184
1.6094	0.1856	0.3584	0.5184	0.6656

(and consequently the mean square error) increased with  $y_1$  and with  $y_2$ . All these facts were expected. The performance of the four estimators for the bivariate distribution function is dominated by the variability of the estimators, with a small advantage for the weighted cumulative hazard estimator and the estimator based on Kaplan–Meier weights, labeled as WCH and KMW, respectively.

**Table 2** Mean square errors ( $\times 100$ ) for the estimators of the bivariate distribution function

$y_2$	$y_1$			
	0.2231	0.5108	0.9163	1.6094
KMW estimator				
0.2231	0.0278	0.0447	0.0603	0.0699
0.5108	0.0467	0.0803	0.1073	0.1316
0.9163	0.0659	0.1049	0.1382	0.1807
1.6094	0.0814	0.1364	0.1744	0.2981
Lin's estimator				
0.2231	0.0330	0.0497	0.0628	0.0659
0.5108	0.0531	0.0895	0.1145	0.1272
0.9163	0.0741	0.1216	0.1608	0.1959
1.6094	0.1057	0.1679	0.2358	0.4443
LDM estimator				
0.2231	0.0280	0.0445	0.0580	0.0627
0.5108	0.0527	0.0847	0.1068	0.1222
0.9163	0.0833	0.1356	0.1611	0.1977
1.6094	0.1030	0.1930	0.2505	0.4453
WCH estimator				
0.2231	0.0281	0.0444	0.0575	0.0621
0.5108	0.0465	0.0781	0.1003	0.1125
0.9163	0.0653	0.1018	0.1321	0.1555
1.6094	0.0810	0.1328	0.1665	0.2409

Gap times generated from Gumbel's bivariate distribution function, with censoring times generated from model  $U[0, 3]$  and a sample size of  $n = 250$

**Table 3** True values of the trivariate distribution of the gap times for the Clayton copula

$y_2$	$y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.1170	0.1358	0.1403	0.1420
0.5108	0.1358	0.1678	0.1767	0.1802
0.9163	0.1403	0.1767	0.1871	0.1912
1.6094	0.1420	0.1802	0.1912	0.1956
$y_1 = 0.5108$				
0.2231	0.1358	0.1678	0.1767	0.1802
0.5108	0.1678	0.2443	0.2744	0.2879
0.9163	0.1767	0.2744	0.3193	0.3412
1.6094	0.1802	0.2879	0.3412	0.3682
$y_1 = 0.9163$				
0.2231	0.1403	0.1767	0.1871	0.1912
0.5108	0.1767	0.2744	0.3193	0.3412
0.9163	0.1871	0.3193	0.3974	0.4420
1.6094	0.1912	0.3412	0.4420	0.5062
$y_1 = 1.6094$				
0.2231	0.1420	0.1802	0.1912	0.1956
0.5108	0.1802	0.2879	0.3412	0.3682
0.9163	0.1912	0.3412	0.4420	0.5062
1.6094	0.1956	0.3682	0.5052	0.6100

### 3.2 Scenario 2: three gap times

In the second scenario, we consider two Archimedean copulas to generate data for a model with three recurrent events (leading to three consecutive gap times): the multivariate Clayton copula and the multivariate Frank copula (Nelsen 2006). In the first setting, the successive gap times  $(Y_1, Y_2, Y_3)$  are simulated according to the trivariate Clayton copula. The trivariate Clayton copula is given by  $C(y_1, y_2, y_3) = [\sum_{i=1}^3 y_i^{-\alpha} - 2]^{-1/\alpha}$ ,  $\alpha \in [-1, \infty[ \setminus 0$ . We consider the Clayton copula with exponential margins with rate parameter 1 and  $\alpha = 2$ . The follow-up time was subjected to right censoring,  $C$ , according to uniform models  $U[0, 3]$  and  $U[0, 6]$ . The first model results in 32% of censoring on the first gap time  $Y_1$ , 55% of censoring on the second gap time  $Y_2$  and 68% of censoring on the third gap time  $Y_3$ . The second model decreases these censoring levels to 17%, 32% and about 46%, respectively. Because of space limitation, we only present the results for the first model.

**Table 4** True values of the trivariate distribution of the gap times for the Frank copula

$y_2$	$y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.0246	0.0417	0.0536	0.0617
0.5108	0.0417	0.0718	0.0930	0.1077
0.9163	0.0536	0.0930	0.1212	0.1411
1.6094	0.0617	0.1077	0.1411	0.1649
$y_1 = 0.5108$				
0.2231	0.0417	0.0718	0.0930	0.1077
0.5108	0.0718	0.1264	0.1666	0.1956
0.9163	0.0930	0.1666	0.2229	0.2645
1.6094	0.1077	0.1956	0.2645	0.3166
$y_1 = 0.9163$				
0.2231	0.0536	0.0930	0.1212	0.1411
0.5108	0.0930	0.1666	0.2229	0.2645
0.9163	0.1212	0.2229	0.3048	0.3683
1.6094	0.1411	0.1645	0.3683	0.4524
$y_1 = 1.6094$				
0.2231	0.0617	0.1077	0.1411	0.1649
0.5108	0.1077	0.1956	0.2645	0.3166
0.9163	0.1411	0.2645	0.3683	0.4524
1.6094	0.1649	0.3166	0.4524	0.5697

In a third setting, the successive gap times  $(Y_1, Y_2, Y_3)$  are simulated according to the the trivariate Frank copula. The trivariate Frank copula is given by  $C(y_1, y_2, y_3) = -\frac{1}{\alpha} \ln \left[ 1 + \frac{\prod_{i=1}^3 (e^{-y_i} - 1)}{(e^{-\alpha} - 1)^2} \right]$  with  $\alpha \in R \setminus 0$ . We consider the Frank copula with exponential margins with rate parameter 1 and  $\alpha = 2$ . Again, censoring was generated according to uniform models  $U[0, 3]$  and  $U[0, 2]$ . The first model results in 25% of censoring on the first gap time  $Y_1$ , 50% of censoring on the second gap time  $Y_2$  and 68% of censoring on the third gap time  $Y_3$ . The second model increases these censoring levels to 43%, 69% and about 80%, respectively.

The true values of  $F_{123}(y_1, y_2, y_3)$  are reported in Tables 3 and 4. Tables 5, 6, 7, 8, 9, 10, 11 and 12 report the mean square error and standard deviations for the four estimators. All four methods have worst performance in the right tail, where the censoring effects are stronger. Results shown in Table 11 suggest that the WCH estimator leads to better results for estimating the trivariate distribution  $F_{123}(y_1, y_2, y_3)$  for higher values of  $y_1$  while neither one seems to be uniformly the best for estimating this quantity for small of mid valued of  $y_1$ . In these cases, the

**Table 5** Mean square error ( $\times 100$ ) for the KMW estimator

$y_2$	$y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.0442	0.0509	0.0576	0.0592
0.5108	0.0491	0.0632	0.0700	0.0712
0.9163	0.0516	0.0664	0.0758	0.0824
1.6094	0.0625	0.0690	0.0737	0.0819
$y_1 = 0.5108$				
0.2231	0.0513	0.0687	0.0693	0.0697
0.5108	0.0602	0.0992	0.1120	0.1198
0.9163	0.0698	0.1094	0.1317	0.1504
1.6094	0.0752	0.1158	0.1564	0.1781
$y_1 = 0.9163$				
0.2231	0.0571	0.0735	0.0739	0.0753
0.5108	0.0632	0.1068	0.1080	0.1530
0.9163	0.0770	0.1324	0.1636	0.2454
1.6094	0.0787	0.1454	0.2472	0.4863
$y_1 = 1.6094$				
0.2231	0.0579	0.0763	0.0768	0.0842
0.5108	0.0704	0.1152	0.1510	0.1809
0.9163	0.0711	0.1676	0.2316	0.3974
1.6094	0.0795	0.1841	0.4675	0.9039

Trivariate Clayton copula with censoring times generated from model  $U[0,3]$  and a sample size of  $n = 250$

KMW estimator seems to be a good alternative. The WCH estimator is among the four alternative methods the one that deals more efficiently at points for which  $y_1, y_2$  and  $y_3$  are higher.

For the second scenario (Frank copula), we show in Fig. 1 the boxplots of the estimates of the trivariate probabilities for eight different points  $(x, y, z)$ , corresponding to combinations of the percentiles 20%, 40%, 60% and 80% of the marginal distributions of the gap times. Results are based on 1000 Monte Carlo replicates for the four estimators, with a sample size of  $n = 250$ . The plots shown in this figure were obtained for the censoring level of  $C \sim U[0, 3]$ . The boxplots shown in this figure reveal some results which agree with our findings reported in the previous scenario (Clayton copula). From these plots, it can be seen that all methods have small biases and confirm the good performance of the proposed estimators. The KMW and WCH methods are the methods with less bias and variability.

**Table 6** Standard deviation for the KMW estimator

$y_2$	$y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.0210	0.0226	0.0240	0.0243
0.5108	0.0222	0.0251	0.0265	0.0267
0.9163	0.0227	0.0258	0.0272	0.0287
1.6094	0.0250	0.0263	0.0275	0.0286
$y_1 = 0.5108$				
0.2231	0.0227	0.0263	0.0262	0.0264
0.5108	0.0245	0.0315	0.0335	0.0346
0.9163	0.0264	0.0331	0.0363	0.0387
1.6094	0.0274	0.0339	0.0396	0.0421
$y_1 = 0.9163$				
0.2231	0.0239	0.0270	0.0272	0.0274
0.5108	0.0251	0.0327	0.0328	0.0390
0.9163	0.0277	0.0364	0.0403	0.0496
1.6094	0.0281	0.0382	0.0498	0.0684
$y_1 = 1.6094$				
0.2231	0.0241	0.0276	0.0277	0.0291
0.5108	0.0266	0.0338	0.0389	0.0425
0.9163	0.0267	0.0410	0.0481	0.0621
1.6094	0.0282	0.0429	0.0667	0.0774

Trivariate Clayton copula with censoring times generated from model U[0,3] and a sample size of  $n = 250$

#### 4 survivalREC structure and functionality

To provide biomedical researchers with an easy-to-use tool for obtaining estimates and corresponding plots of the multivariate distributions in recurrent event data, we developed an R package called `survivalREC`. This software enables users to implement all nonparametric estimators discussed in Sect. 2, including the estimators conditionally on current or past covariates. The package, available at the CRAN repository at <https://cran.r-project.org/web/packages/survivalREC/> comprises 15 functions, which are summarized in Table 13. Briefly, there are two main types of functionalities: (i) to estimate bivariate distribution functions in recurrent events with the `KMWdf`, `LDMdf`, `LINdf`, `WCHdf` and `IPCWdf` functions; and (ii) the corresponding extension for the estimation with three gap times with `KMW3df`, `LDM3df`, `LIN3df` and `WCH3df`. Plots for each method are also displayed using a “`multidf`” object, which can be obtained through the `multidf` function. Finally, the remaining auxiliary functions, `Beran`, `KM`, `KMW` and `NWW`, are included inside the previous functions.

**Table 7** Mean square error ( $\times 100$ ) for the Lin's estimator

$y_2$	$y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.0434	0.0485	0.0563	0.0621
0.5108	0.0544	0.0751	0.0948	0.1055
0.9163	0.0685	0.0866	0.1316	0.1981
1.6094	0.0864	0.1293	0.2168	0.6677
$y_1 = 0.5108$				
0.2231	0.0487	0.0605	0.0595	0.0675
0.5108	0.0648	0.1070	0.1032	0.1348
0.9163	0.0817	0.1114	0.1693	0.3179
1.6094	0.1018	0.1752	0.3643	1.0826
$y_1 = 0.9163$				
0.2231	0.0545	0.0658	0.0662	0.0607
0.5108	0.0676	0.1096	0.1209	0.1623
0.9163	0.0836	0.1217	0.1793	0.5178
1.6094	0.0902	0.2006	0.4194	1.3850
$y_1 = 1.6094$				
0.2231	0.0536	0.0620	0.0753	0.0740
0.5108	0.0701	0.1008	0.1555	0.1703
0.9163	0.0846	0.1596	0.2834	0.2950
1.6094	0.0980	0.2279	0.3990	1.0970

Trivariate Clayton copula with censoring times generated from model  $U[0,3]$  and a sample size of  $n = 250$

### 4.1 Application to bladder cancer study data

Bladder cancer is one of the most common genitourinary malignant disease being more common in men than in women. Prognosis of this disease is, in most cases, related to risk factors that include smoking, family history, frequent bladder infections, and exposure to certain chemicals. Another significant prognostic factor for these patients' overall survival is the presence of a recurrence. In fact, bladder cancer is a disease with a high percentage of patients who have superficial tumors, which tend to recur, but which are generally not fatal.

The bladder cancer study data set (Byar 1980) includes 118 patients that entered the study with superficial bladder tumors. Tumors were removed transurethrally and patients were assigned to one of three treatments (placebo, pyridoxine, or thiotepa) and followed until the end of the study. The time period under consideration is over four years since the entry of the first subject (48 months). Many of these patients had multiple recurrences of tumors during the study, and new tumors were removed at

**Table 8** Standard deviation for the Lin’s estimator. Trivariate Clayton copula with censoring times generated from model U[0,3] and a sample size of  $n = 250$

$y_2$	$y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.0208	0.0220	0.0237	0.0249
0.5108	0.0233	0.0274	0.0308	0.0325
0.9163	0.0261	0.0294	0.0363	0.0446
1.6094	0.0294	0.0359	0.0465	0.0760
$y_1 = 0.5108$				
0.2231	0.0221	0.0245	0.0244	0.0260
0.5108	0.0254	0.0327	0.0322	0.0367
0.9163	0.0286	0.0334	0.0412	0.0564
1.6094	0.0319	0.0418	0.0604	0.0913
$y_1 = 0.9163$				
0.2231	0.0234	0.0257	0.0257	0.0246
0.5108	0.0260	0.0331	0.0348	0.0403
0.9163	0.0289	0.0349	0.0423	0.0717
1.6094	0.0301	0.0448	0.0645	0.0921
$y_1 = 1.6094$				
0.2231	0.0232	0.0249	0.0275	0.0272
0.5108	0.0265	0.0317	0.0393	0.0413
0.9163	0.0291	0.0400	0.0530	0.0506
1.6094	0.0313	0.0478	0.0603	0.0687

each visit. The time between tumor recurrence and death or censoring was recorded for each patient. The maximum observed number of recurrences is 9.

In this subsection, we will use data on the 85 subjects with nonzero follow-up who were assigned to either thiotepa or placebo, with respective sizes of 47 and 38. Summary statistics for the two treatments are given in Table 14. Among the 85 patients, 47 relapsed at least once, among these, 29 had a second recurrence, 22 had a third recurrence and 14 had four or more recurrences (16.5%). Thus, in our study only the first three recurrence times  $T_1, T_2$  and  $T_3$  (or the corresponding gap times  $Y_1, Y_2$  and  $Y_3$ ) are considered. Data sets considering two, three and four recurrences are available in the `survivalREC` package. To illustrate our methods we will use data with only the first three recurrences for any patient. Below, is an excerpt of the data. frame with one row per individual.



**Table 9** Mean square error ( $\times 100$ ) for the LDM estimator

$y_2$	$y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.0471	0.0523	0.0552	0.0644
0.5108	0.0994	0.0975	0.0992	0.0968
0.9163	0.2795	0.2457	0.1777	0.1533
1.6094	0.5425	0.4548	0.3341	0.2402
$y_1 = 0.5108$				
0.2231	0.0500	0.0622	0.0658	0.0722
0.5108	0.0728	0.1101	0.1252	0.1440
0.9163	0.1919	0.2109	0.2390	0.2585
1.6094	0.6387	0.5895	0.6354	0.5870
$y_1 = 0.9163$				
0.2231	0.0561	0.0666	0.0684	0.0685
0.5108	0.0712	0.1067	0.1297	0.1666
0.9163	0.1315	0.1464	0.2198	0.3296
1.6094	0.6115	0.4615	0.6247	1.0611
$y_1 = 1.6094$				
0.2231	0.0538	0.0668	0.0789	0.0755
0.5108	0.0680	0.1054	0.1454	0.1616
0.9163	0.1133	0.1492	0.1967	0.4431
1.6094	0.5136	0.3332	0.6386	1.9065

Trivariate Clayton copula with censoring times generated from model U[0,3] and a sample size of  $n = 250$

```

> library(survivalREC)

> data("bladder4state")

> head(bladder4state)

      id y1 d1 y2 d2 y3 d3 rx size
1  1  1  0  0  0  0  0  1   3
2  2  4  0  0  0  0  0  1   1
3  3  7  0  0  0  0  0  1   1
4  4 10  0  0  0  0  0  1   1
5  5  6  1  4  0  0  0  1   1
6  6 14  0  0  0  0  0  1   1

> dim(bladder4state)
[1] 85  9
    
```

The movement among the recurrent events is given by the variables  $y_i$  and  $d_i$ , with  $i = \{1, \dots, 4\}$ , which represent, respectively, the four gap times and their corresponding censoring indicators (1 for an event and 0 for censoring). The other

**Table 10** Standard deviation for the LDM estimator

$y_2$	$y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.0215	0.0224	0.0233	0.0253
0.5108	0.0235	0.0252	0.0269	0.0281
0.9163	0.0249	0.0241	0.0264	0.0270
1.6094	0.0239	0.0236	0.0236	0.0273
$y_1 = 0.5108$				
0.2231	0.0223	0.0248	0.0255	0.0268
0.5108	0.0264	0.0322	0.0341	0.0355
0.9163	0.0332	0.0346	0.0395	0.0417
1.6094	0.0358	0.0404	0.0466	0.0484
$y_1 = 0.9163$				
0.2231	0.0237	0.0258	0.0262	0.0262
0.5108	0.0265	0.0326	0.0359	0.0404
0.9163	0.0336	0.0345	0.0422	0.0525
1.6094	0.0426	0.0499	0.0580	0.0681
$y_1 = 1.6094$				
0.2231	0.0232	0.0258	0.0281	0.0274
0.5108	0.0259	0.0324	0.0376	0.0390
0.9163	0.0327	0.0373	0.0424	0.0567
1.6094	0.0482	0.0490	0.0627	0.0799

Trivariate Clayton copula with censoring times generated from model  $U[0,3]$  and a sample size of  $n = 250$

three variables are the patient id (“id”), the type of treatment (“rx”, 1 = *placebo* and 2 = *thiotepa*), and the size (*cm*) of the largest initial tumour (“size”).

One important goal of these studies is to evaluate the effect on future prognosis of a locoregional recurrence (LR), since it is well known that the increased risk after a recurrence decreases significantly with increasing time since LR. In bladder cancer studies, a LR is called an early recurrence if the cancer comes back 6 to 12 months after treatment, and a late recurrence otherwise. The curves depicted in the first row of Fig. 2 show the results for the four proposed methods for the bivariate distribution function ( $F_{12}(x, y)$ ) when  $x = 6$  or  $x = 12$  are fixed. With the exception of the LIN estimator, all the remaining three methods shown in these plots report roughly the same estimates. In fact, a specific issue with the LIN estimator is visible at the top-right of these two figures, because the displayed curves are not monotonically decreasing in  $y$ . This is a consequence of the specific reweighting of the data that is used in this approach, which may lead to problems of interpretation at the right tail of the distribution. The remaining plots in Fig. 2 (second and third rows) are intended to demonstrate the behavior of the four different methods for estimating the trivariate distribution ( $F_{123}(x, y, z)$  when different values of  $x, y,$  and  $z$  are used). The analysis of these plots revealed that, besides the LIN method, the LDM method also has the

**Table 11** Mean square error ( $\times 100$ ) for the WCH estimator

$Y_2$	$Y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.0424	0.0465	0.0514	0.0580
0.5108	0.0500	0.0598	0.0696	0.0757
0.9163	0.0559	0.0618	0.0764	0.0832
1.6094	0.0575	0.0687	0.0734	0.0990
$y_1 = 0.5108$				
0.2231	0.0479	0.0587	0.0587	0.0652
0.5108	0.0633	0.0932	0.0951	0.1025
0.9163	0.0701	0.0966	0.1247	0.1432
1.6094	0.0741	0.1143	0.1738	0.2062
$y_1 = 0.9163$				
0.2231	0.0542	0.0642	0.0644	0.0618
0.5108	0.0646	0.0969	0.1126	0.1252
0.9163	0.0697	0.0960	0.1396	0.1795
1.6094	0.0700	0.1364	0.2146	0.3473
$y_1 = 1.6094$				
0.2231	0.0522	0.0620	0.0698	0.0694
0.5108	0.0640	0.0953	0.1259	0.1234
0.9163	0.0724	0.1240	0.1531	0.2427
1.6094	0.0739	0.1433	0.2479	0.4994

Trivariate Clayton copula with censoring times generated from model  $U[0,3]$  and a sample size of  $n = 250$

drawback of occasionally providing estimated curves that are clearly non-monotone for the trivariate distribution function and, therefore, their practical use could be less recommended. The WCH method and the method based on the Kaplan–Meier weights (KMW) both show plausible curves.

The WCH method provides a nice approach that can be used to estimate the conditional distribution function of the second gap time (time to second recurrence) conditional on the first gap time (time to the first recurrence). The plot shown in Fig. 3 depicts the estimates of  $P(Y_2 \leq y | Y_1 \leq 12)$  and  $P(Y_2 \leq y | Y_1 > 12)$  using the WCH method. The estimated curves reveal that patients with a late recurrence (after 12 months) have a reduced risk of developing a second recurrence.

In what follows, we explain how to use the `survivalREC` package to get estimates and plots for the bivariate distribution and for the distribution functions with three gap times. For illustration purposes, we will consider the WCH method, and the first (top left) and last (bottom right) plots shown in Fig. 2. First, to get the corresponding plots for the bivariate distribution function we need to transform the original data set into a “multidf” format class. This can be done using the `multidf` function that has as arguments `time1`, `time`, `event1` and `status`. These

arguments correspond to the sojourn time in the initial state and the global time, as well as their corresponding censoring indicator variables.

```
> b3state<-multidf(gap1=bladder4state$y1, event1=bladder4state$d1,
                  gap2=bladder4state$y2, status=bladder4state$d2,
                  size=bladder4state$size)
> class(b3state)
[1] "multidf"
```

To obtain the nonparametric estimates for bivariate distribution, we have the functions `KMWdf`, `LDMdf`, `LINDf` and `WCHdf`.

As an example, suppose we are interested in obtaining the estimates for  $F_{12}(x = 6, y = 20)$  for the method `WCH`. The input codes for this case are the following:

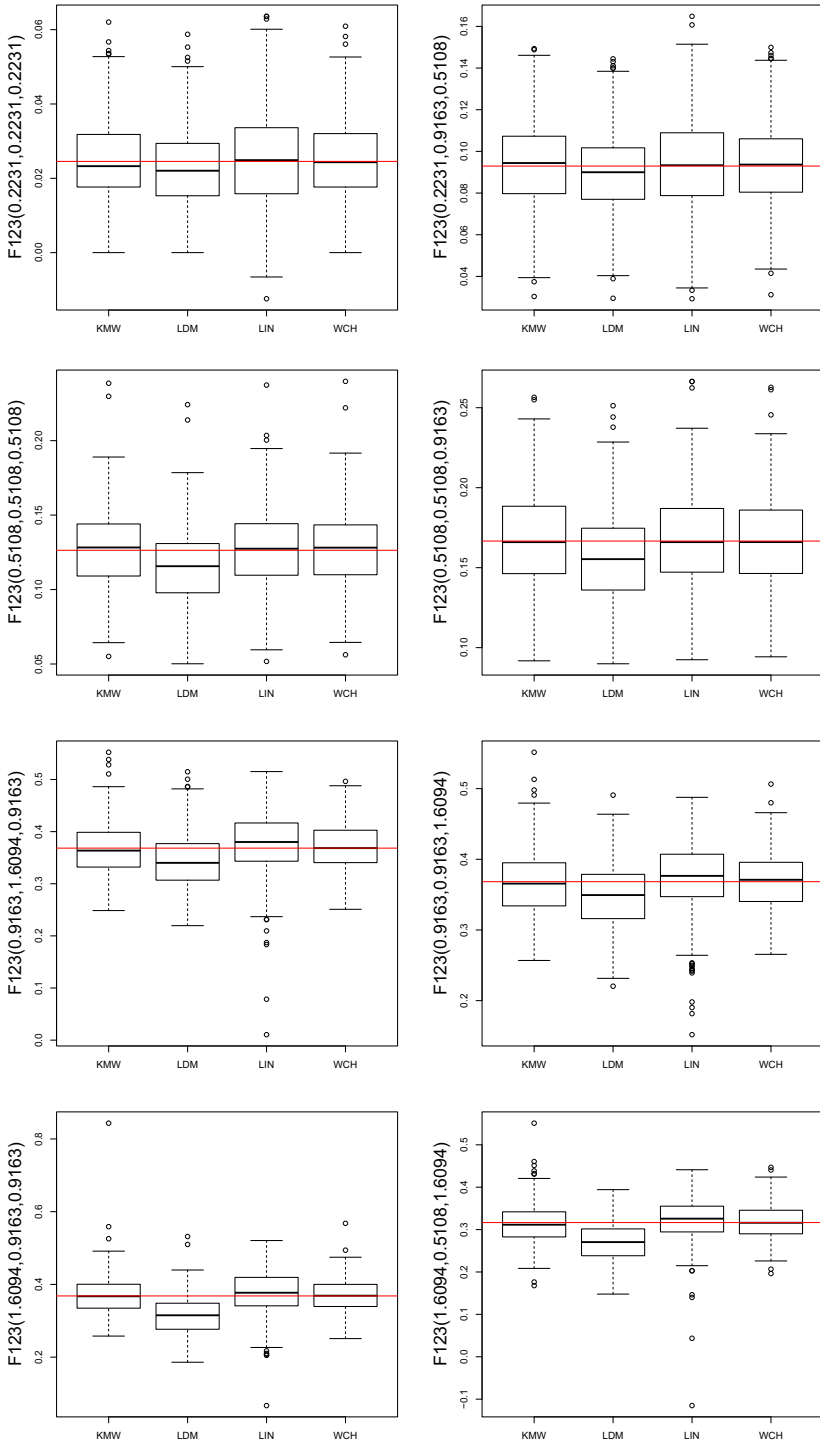
```
> WCHdf(b3state, x=6, y=20)
[1] 0.2078159
```

It is also possible to show the estimated curves of the bivariate distribution function given a specific value for the first gap time. This can be done through the

**Table 12** Standard deviation for the WCH estimator

$y_2$	$y_3$			
	0.2231	0.5108	0.9163	1.6094
$y_1 = 0.2231$				
0.2231	0.0206	0.0216	0.0227	0.0241
0.5108	0.0224	0.0245	0.0264	0.0275
0.9163	0.0237	0.0248	0.0276	0.0289
1.6094	0.0240	0.0262	0.0271	0.0314
$y_1 = 0.5108$				
0.2231	0.0219	0.0242	0.0242	0.0255
0.5108	0.0252	0.0305	0.0309	0.0320
0.9163	0.0264	0.0311	0.0353	0.0379
1.6094	0.0272	0.0337	0.0417	0.0447
$y_1 = 0.9163$				
0.2231	0.0233	0.0253	0.0254	0.0249
0.5108	0.0254	0.0311	0.0336	0.0354
0.9163	0.0264	0.0309	0.0372	0.0424
1.6094	0.0265	0.0370	0.0464	0.0570
$y_1 = 1.6094$				
0.2231	0.0228	0.0249	0.0264	0.0263
0.5108	0.0253	0.0309	0.0352	0.0351
0.9163	0.0269	0.0351	0.0391	0.0493
1.6094	0.0272	0.0378	0.0498	0.0690

Trivariate Clayton copula with censoring times generated from model  $U[0,3]$  and a sample size of  $n = 250$



- ◀ **Fig. 1** Boxplots of the  $M = 1000$  estimates of the trivariate distribution function of the four estimators, with a sample sizes of 250. Data from Frank copula with unit correlated exponential distributions. Censoring times were generated from an uniform distribution on  $[0,3]$ . Horizontal solid red line corresponds to the true value

`plot` function for “multidf” objects. The following input codes show how to obtain the graphic for method WCH in which the first gap time ( $t_1$ ) takes value 6.

```
> plot(x=b3state, t1=6, method="WCH", type = "s", ylab='Prob.',
      ylim=c(0,.4), xlab='F12(x=6,y)')
> legend("bottomright" , legend=c("WCH"), col=c("Black"), lty=1,
      cex=0.8)
```

The procedures to extend the estimation to three gap times are quite similar to the bivariate case. First, we must create a new object using the `multidf` function, called for this example, `b4state`. As we can see, the `b4state` object gives us the cumulative gap times `time1`, `time2` and `time`, as well as, the corresponding censoring indicators (`event1`, `event2` and `status`). Finally, to obtain estimates, the `WCH3df` function can be used by adding a new parameter,  $z$ , to the third gap time.

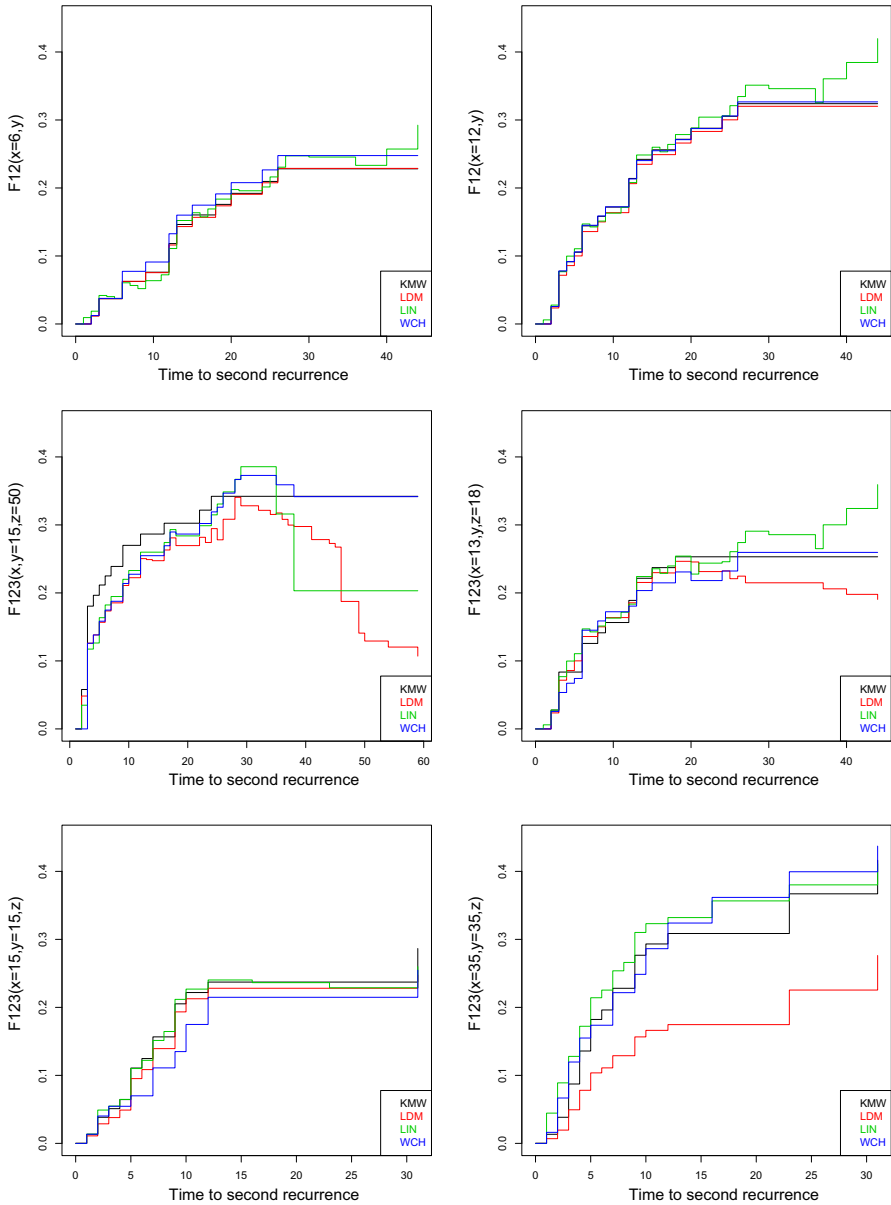
## 5 Conclusions and final remarks

There have been several contributions to the estimation of the marginal and joint distributions in the context of recurrent events. The methods based on the inverse probability of censoring weights introduced by Lin et al. (1999) and the method based on Kaplan–Meier weights (de Uña- Álvarez and Meira-Machado 2008) are among the best ones for estimating the bivariate distribution function.

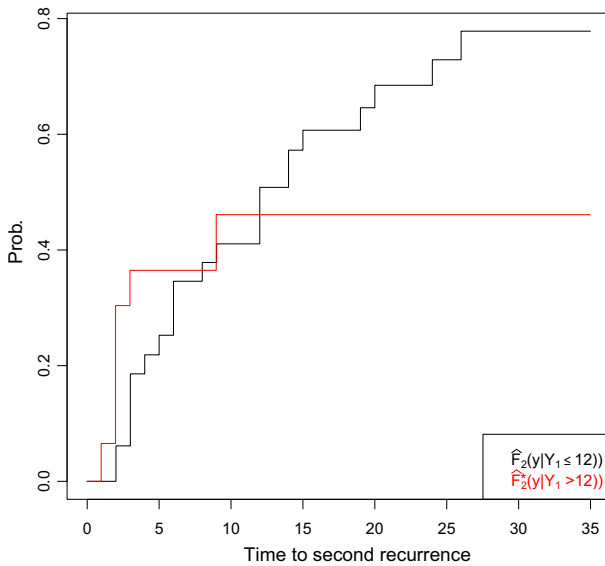
In this paper, we introduce new nonparametric methods for the estimation of these quantities. Simulations show that most of the proposed estimators for the bivariate distribution function are virtually unbiased. The extension of these methods to several gap times is also discussed.

To provide biomedical researchers with an easy-to-use tool, we have also developed the `survivalREC` package, which is available at the CRAN repository, which enables one to implement all proposed methods. Details on the usage of its functions and main functionalities are also introduced in the paper through an illustrative example of the analysis of recurrence events in a bladder cancer study.

Another issue of practical interest that is studied in this paper is the estimation of these quantities conditionally on current or past covariate measures. A feasible nonparametric solution to this problem is proposed. The proposed method is based on local smoothing by the means of kernel weights that are either based on a local constant (i.e., Nadaraya–Watson) or a local linear regression.



**Fig. 2** Estimates of the bivariate (first row) and trivariate (second and third rows) distribution using the four proposed methods. Bladder recurrence cancer data



**Fig. 3** Estimates of the distribution of the second gap time (time to second recurrence) conditional on the time to the first recurrence. Estimated curves based on the WCH approach. Bladder recurrence cancer data

**Table 13** Summary of functions in the `survivalREC` package

Function	Description
<code>multidf</code>	Create a <code>multidf</code> object
<code>KMwdf</code>	Estimation of the bivariate distribution function based on Kaplan–Meier Weights
<code>LDMdf</code>	Estimation of the bivariate distribution function based on landmarking
<code>LINdf</code>	Estimation of the bivariate distribution function using Lin’s estimator
<code>WCHdf</code>	Estimation of the bivariate distribution function based on the Weighted Cumulative Hazard estimator (WCH)
<code>KMW3df</code>	Estimation of distribution with three gap times. Method based on Kaplan–Meier Weights
<code>LDM3df</code>	Estimation of distribution with three gap times. Method based on landmarking
<code>LIN3df</code>	Estimation of distribution with three gap times. Method based on the extension the Lin’s estimator
<code>WCH3df</code>	Estimation of distribution with three gap times. Method based on the extension of the Weighted cumulative hazard estimator (WCH)
<code>IPCWdf</code>	Estimation of the bivariate distribution function based on the Inverse Probability of Censoring Weighting estimator (IPCW)
<code>plot.multidf</code>	Plot for an object of class <code>multidf</code>
<code>Beran</code>	Computes the conditional survival probability of the response, given the covariate under random censoring
<code>KM</code>	Computes the Kaplan–Meier product-limit of survival
<code>KMW</code>	Returns a vector with the Kaplan–Meier weights
<code>NWW</code>	Returns a vector with the Nadaraya–Watson weights



**Table 14** Recurrent experiences for placebo and thiotepa groups in Stage I bladder cancer clinical trial

Treatment	Recurrences						
	0	1	2	3	4	5	> 5
Placebo	18	10	4	6	2	4	3
Thiotepa	20	8	3	2	2	2	1

**Acknowledgements** This research was financed by Portuguese Funds through FCT—“Fundação para a Ciência e a Tecnologia”, within Projects projects UIDB/00013/2020, UIDP/00013/2020 and the research grant PD/BD/142887/2018.

## References

- Amorim L, Cai J (2015) Modelling recurrent events: a tutorial for analysis in epidemiology. *Int J Epidemiol* 44(1):324–333
- Andersen PK, Gill RD (1982) Cox’s regression model for counting processes: a large sample study. *Ann Stat* 10:1100–1120
- Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York
- Beran R (1981) Nonparametric regression with randomly censored survival data. Technical report. Univ. California, Berkeley
- Breslow NE (1972) Discussion of the paper by D. R. Cox. *J R Stat Soc B* 34:216–217
- Burke MD (1988) Estimation of a bivariate distribution function under random censorship. *Biometrika* 75:379–382
- Byar DP (1980) The Veterans Administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine and topical thiotepa. In: Pavone-Macaluso M, Smith PH, Edsmyr F, Plenum (eds) Bladder tumors and other topics in urological oncology. Springer, New York, pp 363–370
- Campbell G (1981) Non-parametric bivariate estimation with randomly censored data. *Biometrika* 68:417–423
- Cook RJ, Lawless JF (2007) The analysis of recurrent event data. Springer, New York
- Cox DR (1972) Regression models and life-tables. *J R Stat Soc B* 34(2):187–220
- de Uña-Álvarez J, Amorim AP (2011) A semiparametric estimator of the bivariate distribution function for censored gap times. *Biom J* 53(1):113–127
- de Uña-Álvarez J, Meira-Machado L (2008) A simple estimator of the bivariate distribution function for censored gap times. *Stat Probab Lett* 78:2440–2445
- 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(2):364–375
- Kaplan E, Meier P (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53(282):457–481
- Lin DY, Ying Z (1993) A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* 80:573–581
- Lin DY, Sun W, Ying Z (1999) Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* 86:59–70
- Meira-Machado L, Faria S (2014) A simulation study comparing modeling approaches in an illness-death multi-state model. *Commun Stat-Simul Comput* 43(5):929–946
- Meira-Machado L, Sestelo M (2016) condSURV: an R package for the estimation of the conditional survival function for ordered multivariate failure time data. *R J* 8(2):460–472
- Meira-Machado L, Sestelo M (2019) Estimation in the progressive illness-death model: a nonexhaustive review. *Biom J* 61(2):245–263
- Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, Andersen P (2009) Multi-state models for the analysis of time to event data. *Stat Methods Med Res* 18:195–222

- Meira-Machado L, de Uña-Álvarez J, Datta S (2015) Nonparametric estimation of conditional transition probabilities in a non-Markov illness-death model. *Comput Stat* 30:377–397
- Meira-Machado L, Sestelo M, Gonçalves A (2016) A simulation study comparing modeling approaches in an illness-death multi-state model. *Biom J* 58(3):623–634
- Moreira A, Meira-Machado L (2012) survivalBIV: estimation of the bivariate distribution function for sequentially ordered events under univariate censoring. *J Stat Softw* 46(13):1–16
- Moreira A, Araújo A, Meira-Machado L (2017) Estimation of the bivariate distribution function for censored gap times. *Commun Stat Simul Comput* 46(1):275–300
- Nadaraya E (1965) On nonparametric estimates of density functions and regression curves. *Theory Appl Probab* 10:186–190
- Nelsen RB (2006) An introduction to copulas. Springer, New York
- Peña EA, Strawderman RL, Hollander M (2001) Nonparametric estimation with recurrent event data. *J Am Stat Assoc* 96:1299–1315
- Pepe MS, Cai J (1993) Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *J Am Stat Assoc* 88:811–820
- Prentice RL, Cai J (1992) Covariance and survival function estimation using censored multivariate failure time data. *Biometrika* 79:495–512
- Prentice RL, Williams BJ, Peterson AV (1981) On the regression analysis of multivariate failure time data. *Biometrika* 68:373–79
- Prentice RL, Moodie FZ, Wu J (2004) Hazard-based nonparametric survivor function estimation. *J R Stat Soc B* 66:305–319
- Schaubel DE, Cai J (2004) Non-parametric estimation of gap-time survival functions for ordered multivariate failure time data. *Stat Med* 23:1885–1900
- Soutinho G, Meira-Machado L (2020) Some of the most common copulas for simulating complex survival data. *Int J Math Comput Simul* 14:28–37
- Tsai WY, Leurgans S, Crowley J (1986) Nonparametric estimation of a bivariate survival function in the presence of censoring. *Ann Stat* 14:1351–1365
- van der Laan MJ, Hubbard AE, Robins JM (2002) Locally efficient estimation of a multivariate survival function in longitudinal studies. *J Am Stat Assoc* 97:494–507
- van Houwelingen HC (2007) Dynamic prediction by landmarking in event history analysis. *Scand J Stat* 34:70–85
- Van Keilegom I (2004) A note on the nonparametric estimation of the bivariate distribution under dependent censoring. *J Nonparametr Stat* 16:659–670
- Wang M-C, Chang S-H (1999) Nonparametric estimation of a recurrent survival function. *J Am Stat Assoc* 94:146–153
- Wang MC, Wells MT (1998) Nonparametric estimation of successive duration times under dependent censoring. *Biometrika* 85:561–572
- Wei LJ, Lin DY, Weissfeld L (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Stat Assoc* 84(408):1065–1073

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.