REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 83–98

INFERENCE FOR NON-MARKOV MULTI-STATE MODELS: AN OVERVIEW

Authors: Luís Meira-Machado
– Department of Mathematics and Applications,
University of Minho, Portugal
lmachado@math.uminho.pt

Abstract:

- In longitudinal studies of disease, patients can experience several events across a follow-up period. Analysis of such studies can be successfully performed by multi-state models. This paper considers nonparametric and semiparametric estimation of important targets in multi-state modeling, such as the transition probabilities and bivariate distribution function (for sequentially ordered events). These estimators are shown to be consistent even for data which is non-Markov. We illustrate the methods on two data sets.

Key-Words:

- *bivariate censoring; Markov property; multi-state models; Kaplan–Meier; presmoothing; transition probabilities.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

In many cancer studies, the main outcome under assessment is the time to death. However, other types of events can be observed during the follow-up period. For example, in colon cancer studies more than one event is often observed such as “local recurrence”, “distant metastasis” and “dead”. The occurrence of these intermediate events often affect patient’s prognosis and can be modeled using a Cox proportional hazards model with a time-dependent covariate. Alternatively, a natural way to model such data is by using a multi-state model with states based on the values of these categorical-valued time-dependent covariates.

A multi-state model is a model for a stochastic process which occupies one of a set of discrete states at any time. These models are well adapted for modeling complex event histories (Andersen *et al.* [1]; Hougaard [2]; Meira-Machado *et al.* [3]). The use of such models is very useful for describing event history data offering a better understanding of the process of the illness, and leading to a better knowledge of the evolution of the disease over time. Issues of interest include the estimation of progression rates, assessing the effects of individual risk factors, survival rates or prognostic forecasting.

The complexity of a multi-state model greatly depends on the number of states defined and by the transitions allowed among these states. The simplest form of multi-state model is the “two-state model”, or mortality model, for survival analysis (with only two states, “Alive” and “Dead”, and a single transition). Splitting the “Alive” state from the simple mortality model for survival data into two transient states, we therefore obtain the simplest progressive three-state model, illustrated in Figure 1. Graphically, multi-state models may be illustrated using diagrams with rectangular boxes to represent possible states and with arrows between the states representing the allowed transitions. States can be transient or absorbing. A state is said to be an absorbing state if no transitions can emerge from the state (e.g. death). Irreversible illness-death models are often used to model disease processes in medical cancer studies.

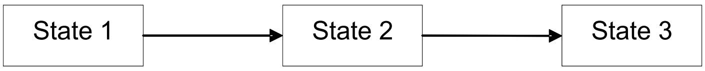

Figure 1: Progressive three-state model.

In these models, individuals may pass from the initial state (e.g. disease-free; state 1), to the intermediate event or disease state (e.g. recurrence; state 2) and then to the absorbing state (e.g. dead; state 3). Individuals are at risk of death in each transient state (states 1 and 2). Figure 2 shows the schematic diagram of transitions involved in the illness-death model.

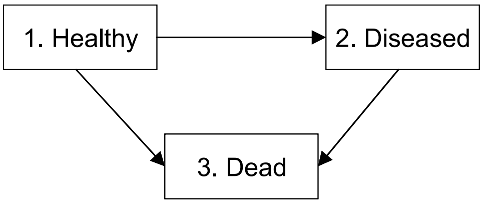

Figure 2: Progressive illness-death model.

The inference in multi-state models is traditionally performed under a Markov assumption for which past and future are independent given its present state (see e.g. [4] and [5]). However, this assumption may fail in some applications, leading to inconsistent estimators. In such cases, alternative (non-Markov) estimators are needed. In this work we review some recent developments in this area, focussing on the estimation of several quantities such as the bivariate distribution function and the transition probabilities. Specifically, we focus on the three-state model of Figure 1 and the illness-death model depicted in Figure 2. In the progressive three-state model, the times between consecutive events (which define states 2 and 3) are often of interest. In Section 2 we present several estimators of the bivariate distribution function of the gap times. Some related problems as estimation of the marginal distribution of the second gap time is discussed. In the framework of the illness-death model, several estimators for the transition probabilities are presented in Section 3. In Section 4, an example of application on bladder tumor recurrence data is re-analyzed to assess the proposed models and methodologies. We also apply our estimation procedures to data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Finally we conclude with a discussion section.

2. ESTIMATION OF THE BIVARIATE DISTRIBUTION

2.1. Notation

Assume the progressive three-state model of Figure 1. Let (T_{12}, T_{23}) 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_{12}, T_{23}) and let $Y = T_{12} + T_{23}$ be the total time. Because of this, we only observe $(\widetilde{T}_{12i}, \widetilde{T}_{23i}, \Delta_{1i}, \Delta_{2i})$, $1 \le i \le n$, which are n independent replications of $(\widetilde{T}_{12}, \widetilde{T}_{23}, \Delta_1, \Delta_2)$, where $\widetilde{T}_{12} = T_{12} \wedge C$, $\Delta_1 = I(T_{12} \le C)$, and $\widetilde{T}_{23} = T_{23} \wedge C_2$, $\Delta_2 = I(T_{23} \le C_2)$ with $C_2 = (C - T_{12})\, I(T_{12} \le C)$ the censoring variable of the second gap time. Since $\Delta_2 = 1$ implies $\Delta_1 = 1$ then $\Delta_2 = \Delta_2 \Delta_1 = I(Y \le C)$ is the censoring indicator pertaining to the total time. Define $\widetilde{Y} = Y \wedge C$ and let F_1 and G denote the distribution functions of T_{12} and C, respectively. Since T_{12} and C are independent, the Kaplan–Meier estimator based on the pairs $(\widetilde{T}_{12i}, \Delta_{1i})$'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 $(\widetilde{T}_{12i} + \widetilde{T}_{23i}, \Delta_{2i})$'s. Because T_{23} 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 $F_{12}(x, y) = P(T_{12} \le x, T_{23} \le y)$. This issue have received much attention recently. Among others it was investigated by Lin *et al.* [6], Van Keilegom [7], de Uña-Álvarez and Meira-Machado [8] or de Uña-Álvarez and Amorim [9].

In this section we present four estimators for the bivariate distribution function of the gap times. All estimator are somehow related since all use (in different ways) the Kaplan–Meier estimator [10].

2.2. Methods

A simple estimator for the bivariate distribution function of the gap times is based on the Kaplan–Meier survival function (Conditional Kaplan–Meier, CKM).

Since $F_{12}(x, y) = P(T_{12} \le x, T_{23} \le y) = P(T_{23} \le y \mid T_{12} \le x)\, P(T_{12} \le x)$ one simple estimator for the bivariate distribution is given by

$$\widehat{F}_{12}(x, y) \;=\; \widehat{F}_1(x)\, \widehat{F}_{\mathrm{KM}}\big(y \mid T_{12} \le x, \Delta_1 = 1\big) \tag{2.1}$$

where $\widehat{F}_1(x)$ is the Kaplan–Meier product-limit estimator based on the pairs $(\widetilde{T}_{12i}, \Delta_{1i})$'s. The $\widehat{F}_{\mathrm{KM}}(y \mid T_{12} \le x, \Delta_1 = 1)$ is the conditional distribution function for the subset of $T_{12} \le x$ and $\Delta_1 = 1$ (the Kaplan–Meier estimator based on the subset $(\widetilde{T}_{23i}, \Delta_{2i})$'s such that $\widetilde{T}_{12i} \le x$ and $\Delta_{1i} = 1$).

Another estimator for the bivariate distribution function was proposed by Lin *et al.* [6]. This estimator is based on inverse probability of censoring weighted (IPCW) and is expressed as

$$\overset{\frown}{F}_{12}(x, y) \;=\; \overset{\frown}{H}(x, 0) - \overset{\frown}{H}(x, y) \tag{2.2}$$

where

$$\overset{\frown}{H}(x, y) \;=\; \frac{1}{n} \sum_{i=1}^{n} \frac{I\big(\widetilde{T}_{12i} \le x, \widetilde{T}_{23i} > y\big)}{1 - \widehat{G}\big((\widetilde{T}_{12i} + y)^-\big)}$$

and where $\widehat{G}$ stands for the Kaplan–Meier estimator of the censoring distribution based on the $(\widetilde{Y}_i, 1-\Delta_{2i})$'s.

Recently de Uña-Álvarez and Meira-Machado [8] proposed a simple estimator for the bivariate distribution. The idea behind the estimator is using the Kaplan–Meier estimator pertaining to the distribution of the total time to weight the bivariate data. The proposed estimator (Weighted Kaplan–Meier, WKM) is given by

$$\widetilde{F}_{12}(x,y) \;=\; \sum_{i=1}^{n} W_i\, I\big(\widetilde{T}_{12i} \le x, \widetilde{T}_{23i} \le y\big) \tag{2.3}$$

where

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

are the Kaplan–Meier weights attached to $\widetilde{Y}_i$ when estimating the marginal distribution of Y from $(\widetilde{Y}_i, \Delta_{2i})$'s, and for which the ranks of the censored $\widetilde{Y}_i$'s, R_i, are higher than those for uncensored values in the case of ties.

An estimator related to (2.3) was recently proposed by de Uña-Álvarez and Amorim [9]. In this estimator they assume a presmoothed version of the Kaplan–Meier estimator (see [11] and [12] for more details). Presmoothing goes back at least to Dikta (1998) and the idea is to replace the censoring indicators by some smooth fit. This smooth can be based on a certain parametric family (yielding a semiparametric estimator) or using a nonparametric binary regression curve. The term "presmoothing" comes from the fact that smoothing is simply used to get a modified version of the Kaplan–Meier weights, but the final estimator is not smooth itself. Throughout this paper we will assume that the probability of censoring for the second gap time, T_{23}, given the (possibly censored) gap times belongs to a parametric family of binary regression curves. Put $m(x,y) = P(\Delta_2 = 1 \mid \widetilde{T}_{12} = x, \widetilde{Y} = y)$, that is, the probability of uncensoring for the total time Y given the observable information on both gap times. Then the new estimator (Smooth Weighted Kaplan–Meier, SWKM) is expressed as

$$\overline{F}_{12}(x,y) \;=\; \sum_{i=1}^{n} W_i^{\star}\, I\big(\widetilde{T}_{12i} \le x, \widetilde{T}_{23i} \le y\big) \tag{2.4}$$

where

$$W_i^{\star} \;=\; \frac{m(\widetilde{T}_{12i}, \widetilde{Y}_i)}{n-R_i+1} \prod_{R_j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{12j}, \widetilde{Y}_j)}{n-R_j+1}\right]$$

are the presmoothed Kaplan–Meier weights where each censoring indicator Δ_{2j} in W_i is replaced by the conditional probability of censoring for the second gap time, given the available information. The m function stands for a (smooth) parametric binary regression model, e.g. logistic. In practice, we assume that

$m(x,y) = m(x,y;\beta)$ where β is a vector of parameters which typically will be computed by maximizing the conditional likelihood of the Δ_2's given $(\widetilde{T}_{12}, \widetilde{T}_{23})$ for those with $\Delta_1 = 1$. Thus, we introduce the parametrically presmoothed Kaplan–Meier weights as

$$W_i^{\star}(\beta) \;=\; \frac{m(\widetilde{T}_{12i}, \widetilde{Y}_i; \beta)}{n - R_i + 1} \prod_{R_j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{12j}, \widetilde{Y}_j; \beta)}{n - R_j + 1} \right].$$

Note that, unlike (2.3), the SWKM estimator may attach positive mass to pair of gap times with censored second gap time; but only for those with uncensored first gap time. Conditions under which both estimators are consistent is fully discussed in papers by de Uña-Álvarez and Meira-Machado [8] and de Uña-Álvarez and Amorim [9]. Note that without presmoothing, the estimator (2.4) reduces to (2.3). Without censoring both reduce to the empirical estimator.

It is also important to mention that estimators (2.2), (2.3) and (2.4) are only estimable on $\{(x,y) \colon x + y \leq C_{\max}\}$ where $C_{\max}$ is the maximum follow-up time. This means that consistency of these estimators is only guaranteed on the triangle shown in Figure 3.

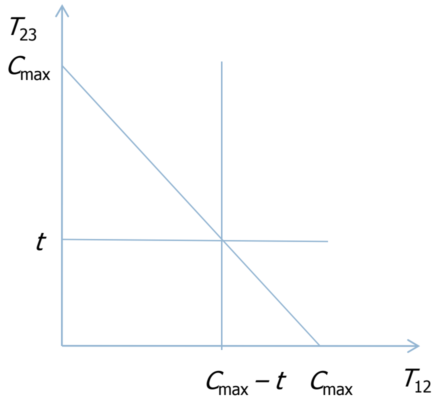

Figure 3: Estimable area of estimators (2.2), (2.3) and (2.4).

We note that the estimates produced via Kaplan–Meier (CKM) may not produce a valid bivariate distribution since it does not guarantee that the bivariate distribution function is monotone. The problem can be explained to the fact that, as the conditioning set $T_{12} \leq x$ changes, the redistribution to the right of the probability mass associated with censored observations also changes. In contrast to the other two methods, the estimators by de Uña-Álvarez and Meira-Machado [8] and de Uña-Álvarez and Amorim [9] are a proper distribution function, in the sense that it attaches positive mass to each observation.

Results of an extensive simulation study comparing the four methods are reported in Meira-Machado and Moreira [13]. The main conclusions are the following:

(**a**) the CKM estimator has larger bias for higher values of the first gap time, but in some cases is one of the estimators with less variance;

(**b**) the WKM estimator has less bias than its smooth version (SWKM); however as expected the later obtained less variance (and less mean square error);

(**c**) the WKM and IPCW estimator are almost unbiased but the last one obtains higher levels of variance for small values of the second gap time.

From the introduced estimators we can obtain an estimator for the marginal distribution of the second gap time, $F_2(y) = P(T_{23} \le y)$, namely

$$\widehat{F}_2(y) \;=\; \widehat{F}_{12}(+\infty, y) \;=\; \widehat{F}_{\mathrm{KM}}(y \,|\, \Delta_1 = 1) \;, \tag{2.5}$$

$$\widetilde{F}_2(y) \;=\; \widetilde{F}_{12}(+\infty, y) \;=\; \sum_{i=1}^{n} W_i \, I(\widetilde{T}_{23i} \le y) \;. \tag{2.6}$$

Note that estimator (2.5), obtained from the CKM, is the Kaplan–Meier estimator based on $(\widetilde{T}_{23i}, \Delta_{2i})$'s such that $\Delta_1 = 1$ (i.e., for which the first gap time is uncensored). Estimator (2.6) is different because the Kaplan–Meier weights W_i in this estimator are based on the $\widetilde{Y}_i$-ranks rather than on the $\widetilde{T}_{23i}$-ranks. Indeed, since T_{23} and C_2 are expected to be dependent, the ordinary Kaplan–Meier estimator of F_2 (estimator (2.5)) will be in general inconsistent. The corresponding estimators for (2.2) and (2.4) are obtained using the same ideas.

2.3. Alternative estimators based on the location-scale model

Other estimators were proposed to estimate the bivariate distribution function. A valid estimator of the bivariate distribution function was provided by Van Keilegom [7] which is based on Akritas [14] estimator. However this approach has some limitations since some smoothing is required. Alternative estimators for the above quantities were also given in Van Keilegom *et al.* [15]. This methodology assumes that the vector of gap times (T_{12}, T_{23}) satisfies the nonparametric location-scale regression model $T_{23} = m(T_{12}) + \sigma(T_{12})\varepsilon$, where the functions m and σ are "smooth", and ε is independent of T_{12}. On the basis of the idea of transfer of tail information, the estimator of the error distribution is used to introduce nonparametric estimators for the bivariate distribution function. As shown by the authors, these estimators will be more efficient than the previous, since it allows for the transfer of tail information from lightly censored areas to heavily ones. More details about these methods can be found in the independent paper by Van Keilegom *et al.* [15].

3. ESTIMATION OF THE TRANSITION PROBABILITIES

3.1. Notation

One major goal in longitudinal multi-state studies is the estimation of transition probabilities. Traditionally these quantities are estimated via a nonparametric model (using e.g. the Aalen–Johansen estimator [4]). In a recent paper, Meira-Machado *et al.* [16] introduce a substitute for the Aalen–Johansen estimator in the case of a non-Markov illness-death model. They showed that the new estimator may behave much more efficiently than the Aalen–Johansen when the Markov assumption does not hold. More recently, Amorim *et al.* [17] propose a modification of Meira-Machado *et al.* [16] estimator based on presmoothing ideas which allows for a variance reduction in the presence of censoring. These estimators will be presented in this section, assuming an illness-death model.

In this section we consider the illness-death model depicted in Figure 2 and we assume that all subjects are in state 1 (‘healthy’) at time $t = 0$. The illness-death model is fully characterized by three transitions: two competing transitions leaving state 1 and one transition to the absorbing ‘dead’ state for those subjects visiting state 2. Therefore, we have three potential transition times, T_{hj}, from state h to state j. This means that a subject not visiting state 2 will reach the absorbing state at time T_{13}, while this time will be $T_{12} + T_{23}$ if the subject passes through state 2 before. We denote by $\rho = I(T_{12} \leq T_{13})$ the indicator of visiting state 2 at some time. Let $Z = T_{12} \wedge T_{13}$ be the sojourn time in state 1, and let $Y = T_{12} + \rho T_{23}$ be the total survival time of the process. In practice, several issues influence the observation of these variables T_{hj}. Whenever $T_{13} \leq T_{12}$, one gets a right censored value of T_{12} and no information on T_{23} is available. Similarly, the value of T_{13} will be censored for those individuals entering state 2. Further, right censoring may appear due to time limitation in following-up or to other causes. This extra censoring is modeling by considering a censoring variable C which is assumed to be independent of the process; finally, we put $\widetilde{Z} = Z \wedge C$ and $\widetilde{Y} = Y \wedge C$ for the censored versions of Z and Y, and $\Delta_1 = I(Z \leq C)$ and $\Delta_2 = I(Y \leq C)$ for the respective censoring indicators.

3.2. Estimators based on the Kaplan–Meier weights

Meira-Machado *et al.* [16] derived estimators for the transition probabilities $p_{11}(s,t)$, $p_{12}(s,t)$, $p_{22}(s,t)$, for a general non-Markov illness-death process without recovery as follows. Let H denote the survival function for Z then the transition probabilities are written as

$$p_{11}(s,t) \;=\; \frac{P(Z>t)}{P(Z>s)} \;=\; \frac{H(t)}{H(s)}\;, \tag{3.1}$$

$$p_{12}(s,t) \;=\; \frac{P(s<Z\le t<Y)}{P(Z>s)} \;=\; \frac{E\big[\varphi_{st}(Z,Y)\big]}{H(s)}\;, \tag{3.2}$$

$$p_{22}(s,t) \;=\; \frac{P(Z\le s,\, t<Y)}{P(Z\le s<Y)} \;=\; \frac{E\big[\widetilde{\varphi}_{st}(Z,Y)\big]}{E\big[\widetilde{\varphi}_{ss}(Z,Y)\big]}\;, \tag{3.3}$$

where $\varphi_{st}(u,v) = I(s<u\le t,\, v>t)$ and $\widetilde{\varphi}_{st}(u,v) = I(u\le s,\, v>t)$.

Then, (3.1) and the denominator of (3.2) only involve the Z variable, and they can be estimated by the ordinary Kaplan–Meier estimator, $\widehat{H}$, based on the pairs $(\widetilde{Z}_i, \Delta_{1i})$'s. The transition probability (3.3) and the numerator of the (3.2) involve expectations of particular transformations of the pair (Z,Y) that can be estimated in different ways. In this section we present two methods to empirically approximate these expectations from the data $\big\{\big(\widetilde{Z}_i, \widetilde{Y}_i, \Delta_{1i}, \Delta_{2i}, \Delta_{1i}\rho_i\big), 1\le i\le n\big\}$, which are assumed to form a random sample of the vector $(\widetilde{Z}, \widetilde{Y}, \Delta_1, \Delta_2, \Delta_1\rho)$.

In Meira-Machado *et al.* [16], the expectations $E\big(\varphi_{st}(Z,Y)\big)$ and $E\big(\widetilde{\varphi}_{st}(Z,Y)\big)$ were estimated by Kaplan–Meier integrals of the form

$$\sum_{i=1}^{n} W_i\, \varphi_{st}(\widetilde{Z}_i, \widetilde{Y}_i)$$

where W_i are the Kaplan–Meier weight attached to $\widetilde{Y}_i$ when estimating the marginal distribution of Y from the $(\widetilde{Y}_i, \Delta_{2i})$'s.

Note that, without right-censoring, the estimator of the transition probabilities reduces to the relative frequency of processes in state j at time t among those in state h at time $s<t$. Meira-Machado *et al.* [16] derived large sample properties of these estimators which may be generalized to more complicated non-Markov processes.

The main weakness of this method [16] is that it provides large standard errors in estimation, specially when there is a large proportion of censored data. In order to overcome this issue Amorim *et al.* [17] propose a modification of Meira-Machado *et al.* (2006)'s estimator based on presmoothing ideas, in the presence of censoring. The implementation of these ideas is straightforward in the case of the progressive three-state model (see Section 2) but not so simple for the illness-death model (as explained below).

In the presmoothed version [17], the expectations in (3.2) and (3.3) are estimated by

$$\sum_{i=1}^{n} W_i^{\star}\, \varphi_{st}(\widetilde{Z}_i, \widetilde{Y}_i)$$

where

$$W_i^\star \;=\; \frac{m(\widetilde{Z}_i,\widetilde{Y}_i)}{n-R_i+1}\prod_{R_j=1}^{i-1}\left[1-\frac{m(\widetilde{Z}_j,\widetilde{Y}_j)}{n-R_j+1}\right]$$

and where $m(z,t)$ stands for an estimator of the binary regression function

$$m(z,t) \;=\; P\bigl(\Delta_2=1\,|\,\widetilde{Z}=z,\widetilde{Y}=t\bigr)\,.$$

The problem in the illness-death model is that the function $m(z,t)$ will typically be discontinuous along the line $t=z$, that is, for those values $(\widetilde{Z},\widetilde{Y})$ corresponding to subjects who are censored while being in state 1 or who suffer a direct transition to the absorbing state. To construct $m(z,t)$ the authors propose to estimate independently two functions: $m_1(z,t)$ such that $m_1(\widetilde{Z},\widetilde{Y})$ is the conditional probability of censoring on Y given $(\widetilde{Z},\widetilde{Y})$ and given that a transition to state 2 is observed; and $m_2(t)$ which is the conditional probability of observing a direct transition from state 1 to state 3 given $\widetilde{Z}=t$ (or $\widetilde{Y}=t$) and given that a transition to state 2 is never observed. These functions can be fitted by some smooth models, so we finally have

$$m(z,t) \;=\; m_1(z,t)\,I(z<t)+m_2(t)\,I(z=t)\;.$$

The estimator $m_1(z,t)$ is based on the subsample $\{i:\Delta_{1i}\rho_i=1\}$, while $m_2(t)$ is computed from $\{i:\Delta_{1i}\rho_i=0\}$. The only condition which is assumed on these two functions is that they should approximate well their targets in a uniform sense (see [17] for more details).

Results from a simulation study comparing the two methods is reported in Amorim *et al.* [17], revealing that the semiparametric estimator is more efficient.

4. EXAMPLES OF APPLICATION

The methods described in Section 2 and Section 3 are illustrated through two real data sets. First, we use data from a bladder cancer study (Byar (1980)) conducted by the Veterans Administration Cooperative Urological Research Group. In addition to this data set we also use the well-known and widely studied colon cancer database. In both data sets, a nonfatal event (recurrence) is observed during the disease course. Also, in both data sets, recurrence is a time-dependent covariate that can be re-expressed as a multi-state model, with states based on the values of the covariate. In the first database all deceased patients died after having a recurrence making it possible for the progressive three-state model to be used (Figure 1). In the second database some subjects died without having a recurrence, making feasible for the illness-death model, depicted in Figure 2, to be used.

4.1. Bladder cancer data

In this study, patients had superficial bladder tumors that were removed by transurethral resection. Many patients had multiple recurrences (up to a maximum of 9) of tumors during the study, and new tumors were removed at each visit. For illustration purposes we re-analyze data from 85 individuals in the placebo and thiotepa treatment groups; these data are available as part of the R survival package. Here, only the first two recurrence times and the corresponding gap times T_{12} and T_{23} are considered. From the total of 85 patients, 47 relapsed at least once and, among these, 29 experienced a new recurrence. We have a total amount of censoring of 66% from which 44.7% is obtained from censored observations on the first gap time. We have about 38% of censored Y's among the uncensored first gap time.

We computed the estimated values for all the estimators of the bivariate distribution function, $F_{12}(x, y)$, introduced in Section 2, for x equal to 3, 13, 29 and 49 and y values 3, 10, 17.75 and 36.75, corresponding to marginal survival probabilities of 0.25, 0.5, 0.75 and 0.95. The estimated values of $F_{12}(x, y)$ are reported in Table 1. In this case it is clearly seen that the four methods can provide quite different results, specially at the right tail of the bivariate distribution, where the censoring effects are stronger.

Table 1: Estimated values of the bivariate distribution function $F_{12}(x, y)$ for different pairs of values. Bladder cancer data.

x	Estimator	y			
		3	10	17.75	36.75
3	CKM	0.0364	0.0607	0.1261	0.1746
	IPCW	0.0320	0.0432	0.1240	0.1726
	WKM	0.0128	0.0427	0.1045	0.1167
	SWKM	0.0328	0.0556	0.1089	0.1203
13	CKM	0.0763	0.1684	0.2533	0.3284
	IPCW	0.0668	0.1510	0.2540	0.3154
	WKM	0.1036	0.1742	0.2511	0.2633
	SWKM	0.1193	0.1814	0.2565	0.2679
29	CKM	0.1513	0.2703	0.3680	0.4499
	IPCW	0.1677	0.2902	0.3830	0.4932
	WKM	0.1729	0.2436	0.3205	0.3482
	SWKM	0.2331	0.2952	0.3704	0.3913
49	CKM	0.1571	0.2801	0.3803	0.4764
	IPCW	0.1556	0.2336	0.4355	0.5457
	WKM	0.2294	0.3001	0.3932	0.4209
	SWKM	0.2652	0.3273	0.4109	0.4318

4.2. Colon cancer data

For illustration, we apply the proposed methods of Section 3 to data from a large clinical trial on patients affected by colon cancer. All subjects underwent a curative surgery for colo-rectal cancer. Unfortunately, some of these patients have residual cancer, which lead to disease recurrence and death (in some cases). From the total of 929 patients, 468 (about 50%) developed recurrence and among these 414 (88%) died. Only 38 patients died without recurrence. The rest of the patients (423) remained alive and disease-free up to the end of the follow-up. The presence of patients that experienced a direct transition from the initial state to the absorbing state leads to the need of using the illness-death model with states "Alive and disease-free" (State 1), "Alive with recurrence" (State 2) and "dead" (State 3). Using Cox proportional hazards models, we verified that the transition rate from state 2 to state 3 is affected by the time spent in the previous state [19]. This allowed us to conclude that the Markov assumption may be unsatisfactory for the colon cancer data set. Note that both methods presented in Section 3 do not make use of the Markov information. We will present estimated transition probabilities calculated using these two approaches.

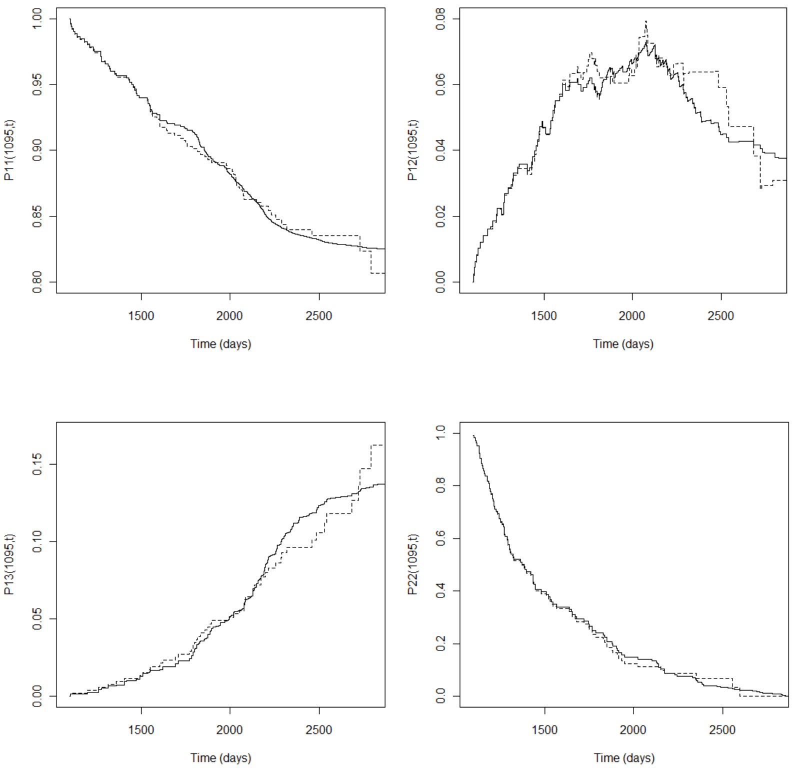

Figure 4: Estimated transition probabilities for $p_{hj}(s,t)$ with $s = 1095$ based on the Kaplan–Meier weights (dashed line) and based on presmoothed Kaplan–Meier weights (solid line).

In Figure 4 we illustrate differences between the estimated transition probabilities, $p_{hj}(s,t)$, $1 \leq h \leq j \leq 3$, based on presmoothing the Kaplan–Meier weights (semiparametric) and the estimator corresponding to no presmoothing [16]. The semiparametric estimator was obtained using a standard logistic model for the parametric estimation of m. The value s was chosen to be as 3 years (1095 days). From this figure we see that the semiparametric estimator have more jump points but with smaller steps. The additional jump points correspond to patients with censored values of the total time that underwent a transition from state 1 to state 2 before time s (uncensored sojourn time in state 1). The number of jump points and the size of the steps are strictly related to the amount of censoring and to the sample size. As expected, both methods provide similar point estimates at small time values while some departures are appreciated for higher time values. In sum, the semiparametric approach provides more reliable curves with less variability, specially in the right tail.

5. DISCUSSION

In this paper we present nonparametric and semiparametric estimators for quantities of interest in multi-state survival modeling. The interest is focused on the estimation of the bivariate distribution function for censored gap times and the estimation of transition probabilities. For both quantities we present two methods based on the Kaplan–Meier estimator pertaining to the distribution of the total time to weight the data. One of these methods is based on presmoothing the Kaplan–Meier estimator. For this, we assume that the probability of censoring for total time given the (possibly censored) gap times belongs to a parametric family of binary regression curves. Some of these estimators may behave much more efficiently than the competing ones. These methods are illustrated using data from two cancer studies.

ACKNOWLEDGMENTS

Thanks to a referee for careful reading of the paper. Luís F. Meira-Machado acknowledges financial support by Grant PTDC/MAT/104879/2008 (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education. The research was also partially funded by CMAT and FCT under the POCI 2010 program.

REFERENCES

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

[2] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer-Verlag, New York.

[3] Meira-Machado, L.; de Uña-Álvarez, J.; Cadarso-Suárez, C. and Andersen, P.K. (2009). Multi-state models for the analysis of time to event data, *Statistical Methods in Medical Research*, **18**, 195–222.

[4] Aalen, O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations, *Scandinavian Journal of Statistics*, **5**, 141–150.

[5] Andersen, P.K. and Keiding, N. (2002). Multi-state models for event history analysis, *Statistical Methods in Medical Research*, **11**, 91–115.

[6] 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.

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

[9] de Uña-Álvarez, J. and Amorim, A.P. (2011). A semiparametric estimator of the bivariate distribution function for censored gap times, *Biometrical Journal*, **53**(1), 113–127.

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

[11] de Uña-Álvarez, J. and Rodríguez-Campos, C. (2004). Strong consistency of presmoothed Kaplan–Meier integrals when covariables are present, *Statistics*, **38**, 483–496.

[12] Dikta, G. (1998). On semiparametric random censorship models, *Journal of Statistical Planning and Inference*, **66**, 253–279.

[13] Meira-Machado, L. and Moreira, A. (2010). Estimation of the bivariate distribution function for censored gap times, *Proceedings of the 19th International Conference on Computational Statistics*, 1367–1374.

[14] Akritas, M.G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring, *The Annals of Statistics*, **22**, 1299–1327.

[15] Van Keilegom, I.; de Uña-Álvarez, J. and Meira-Machado, L. (2011). Nonparametric location-scale models for censored successive survival times, *Journal of Statistical Planning and Inference*, **141**(3), 1118–1131.

[16] Meira-Machado, L.; de Uña-Álvarez, J. and Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness-death model, *Lifetime Data Analysis*, **12**, 325–344.

[17] Amorim, A.P.; de Uña-Álvarez, J. and Meira-Machado, L. (2011). Presmoothing the transition probabilities in the illness-death model, *Statistics & Probability Letters*, *in Press.*

[18] Byar, D.P. (1980). Veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine and topical thiotepa, *Bladder Tumors and Other Topics in Urological Oncology*, **18**(36), 363–370.

[19] Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies, *Biometrics*, **42**, 855–865.