TPmsm: Estimation of the Transition Probabilities in 3-State Models

Artur Araújo
University of Minho

Luís Meira-Machado
University of Minho

Javier Roca-Pardiñas
University of Vigo

Abstract

One major goal in clinical applications of multi-state models is the estimation of transition probabilities. The usual nonparametric estimator of the transition matrix for non-homogeneous Markov processes is the Aalen-Johansen estimator (Aalen and Johansen 1978). However, two problems may arise from using this estimator: first, its standard error may be large in heavy censored scenarios; second, the estimator may be inconsistent if the process is non-Markovian. The development of the R package **TPmsm** has been motivated by several recent contributions that account for these estimation problems. Estimation and statistical inference for transition probabilities can be performed using **TPmsm**. The **TPmsm** package provides seven different approaches to three-state illness-death modeling. In two of these approaches the transition probabilities are estimated conditionally on current or past covariate measures. Two real data examples are included for illustration of software usage.

Keywords: survival, Kaplan-Meier, multi-state model, illness-death model, transition probabilities.

A version of this manuscript has been published online in the *Journal of Statistical Software*, on Dec. 2014, with doi:10.18637/jss.v062.i04.

1. Introduction

In many longitudinal studies it is often of interest to investigate time to a certain event. In medicine the event is an ultimate outcome, such as diagnosis of "death" of the patient or "relapse of the disease". In addition to this primary event of interest one may observe also a number of intermediate ("transient") states, such as "local recurrence" and "distant metastasis" in cancer studies. Analysis of such studies where individuals may experience several events can be successfully performed using a multi-state model (MSM). An MSM is a stochastic process $(X(t), t \in T)$ with a finite state space, where $X(t)$ represents the state occupied by the process at time $t \geq 0$. Graphically, these models are represented by diagrams with rectangular boxes and arrows between them indicating respectively possible states and possible transitions. In general, the future state transitions of an MSM may depend on past events. However, for the special case of a Markov model the past and future are independent given its present state. There exists an extensive literature on MSMs. Main contributions include books by Andersen, Borgan, Gill, and Keiding (1993) and Hougaard (2000). Recent reviews on this topic may be found in the papers by Hougaard (1999), Commenges (1999), Andersen and Keiding (2002), Putter, Fiocco, and Geskus (2007) and Meira-Machado, de Uña-Álvarez, Cadarso-Suárez, and Andersen (2009).

The simplest form of an MSM is the mortality model for survival analysis with only two states "alive" and "dead" with a single transition. Other common models include the progressive three-state model, the illness-death model and the competing risks model. The illness-death model is probably the most used model in the literature, in particular for studying progression of many diseases. This model describes the dynamics of healthy subjects who may move to an intermediate "diseased" state before entering into a terminal absorbing state. Many longitudinal medical data with multiple endpoints can be reduced to this structure. Thus, methods developed for the illness-death model have a wide range of applications. There are several issues of interest in an illness-death multi-state model: study of the relationship between covariates and disease evolution; estimation of transition probabilities; state occupation probabilities; distributions of time spent in each state, among other topics. In this paper we will focus on the inference for the transition probabilities. These quantities provide estimates of the clinical prognosis of a patient at a given point in disease progression, allowing long-term predictions of the process.

Aalen and Johansen (1978) introduced a nonparametric estimator for these quantities for non-homogeneous Markov models. Their estimation method extends the Kaplan-Meier estimator (Kaplan and Meier 1958) to Markov chains. As for the Kaplan-Meier estimator, the standard error of the Aalen-Johansen estimator may be large when the censoring is heavy, particularly with a small sample size. To overcome this problem, Moreira, de Uña-Álvarez, and Meira-Machado (2013) propose a modification of Aalen-Johansen estimator based on presmoothing (Dikta 1998), which allows for a variance reduction in the presence of censoring. In a recent paper, Meira-Machado, de Uña-Álvarez, and Cadarso-Suárez (2006) introduce a substitute for the Aalen-Johansen estimator in the case of a non-Markov illness-death model. They show that when the Markov assumption does not hold, their estimator may behave much better than the Aalen-Johansen which may be systematically biased. The idea behind their estimator is to weight the data by the Kaplan-Meier weights pertaining to the distribution of the total survival time of the process. However, by removing the Markov condition, the proposed substitute for the Aalen-Johansen estimator provides undesirably large standard errors. This problem becomes worse when there is a large proportion of censored data. In order to overcome this problem, Amorim, de Uña-Álvarez, and Meira-Machado (2011) propose a modification of the Meira-Machado estimator based on presmoothing. Other estimators were proposed to estimate the transition probabilities. A valid estimator was provided by Keilegom, de Uña-Álvarez, and Meira-Machado (2011) for a progressive three-state model. This methodology assumes that the vector of gap times (time in State 1 and time in State 2) satisfies the nonparametric location-scale regression model, allowing for the transfer of tail information from lightly censored areas to heavily ones. All these approaches assume independent censoring and do not account for the influence of covariates. To this regard in a recent work, in a regression setup, Meira-Machado, de Uña-Álvarez, and Somnath (2012) introduce feasible estimation methods for the transition probabilities in an illness-death model conditionally on current or past covariate measures.

Software for multi-state survival analysis has been developed recently. A comprehensive list of the available packages at the Comprehensive R Archive Network (CRAN) can be seen in the CRAN task view "Survival Analysis" (Allignol and Latouche 2014). An issue of the *Journal of Statistical Software*, entirely devoted to these models, was published in 2011 (Putter 2011). In R (R Core Team 2014) several packages provide functions for estimating the transition probabilities. The **timereg** package (Scheike and Martinussen 2006; Scheike and Zhang 2011)

can be used to obtain the cumulative incidence probability of a specific cause of failure in competing risks data. It also provides an estimate of its variance at each fixed time point, and constructs $(1-\alpha)100\%$ simultaneous confidence bands over a given time interval. The package **cmprsk** (Gray 2014) can also be used to obtain the same quantities. The package **etm** (Allignol, Schumacher, and Beyersmann 2011) computes and displays the transition probabilities for the Aalen-Johansen estimator. This package also features a Greenwood-type estimator of the covariance matrix. The package **msm** (Jackson 2011) can be used to obtain estimates for the transition probabilities in time-homogeneous Markov models. The package **p3state.msm** (Meira-Machado and Roca-Pardiñas 2011) enables the user to perform inference in the illness-death model. The main feature of the package is its ability for obtaining non-Markov estimates for the transition probabilities. Finally, the **msSurv** package (Ferguson, Datta, and Brock 2012) can be used to estimate the state occupation probabilities along with the corresponding variance estimates, and lower and upper confidence intervals. All of the existing software presents, however, some limitations in practice. Most software assumes the process to be Markovian and assumes independent censoring. Furthermore such software does not account for the influence of covariates. In addition, a comparison between different packages is rather difficult because each of the current programs requests its own data structure.

This paper describes the R package **TPmsm** (Araújo, Roca-Pardiñas, and Meira-Machado 2014) which is available from CRAN at `https://CRAN.R-project.org/package=TPmsm`. The package aims at implementing nonparametric and semiparametric estimators for the transition probabilities in 3-state models. The package provides the so-called Aalen-Johansen estimator typically assumed in Markov processes but it also covers alternative methods which have been proved to be consistent even without the Markov assumption. Inverse censoring probability reweighting is used in some methods to deal with right censoring. These approaches lead to consistent estimators even in the presence of dependent censoring. Finally, two different estimators are implemented that account for the influence of covariates. Bootstrap confidence bands are provided for all methods. In this article we explain and illustrate how numerical and graphical output for all methods can be obtained using the **TPmsm** package.

In Section 2 we introduce the notation for the illness-death stochastic model and describe in detail the proposed estimation methods. In Section 3 we describe the implementation of the methods in package **TPmsm**. Some of the methods are illustrated using generated data in Section 4. Finally, Section 5 illustrates the package's capabilities using two real data examples, and Section 6 gives some concluding remarks and proposals for future work.

2. Methodological background

In this paper we consider the progressive illness-death model depicted in Figure 1. We assume that all subjects are in State 1 at time $t = 0$, and that they may either visit State 2 at some time point; or not, going directly to the absorbing state (State 3). The stochastic behavior of the process is represented by a random vector (T_{12}, T_{13}, T_{23}), where T_{hj} is the potential transition from State h to State j, $1 \leq h < j \leq 3$, in which T_{23} is the sojourn time in State 2. In this model we have two competing transitions $1 \to 2$ and $1 \to 3$. Let the sojourn time in State 1 be denoted by $Z = \min(T_{12}, T_{13})$. The survival time of the process is given by $T = I(T_{12} \leq T_{13})(T_{12} + T_{23}) + I(T_{12} > T_{13})T_{13}$. However, censoring may occur due to follow-up

limitations, lost cases and so on. Because of censoring, one observes $(\widetilde{Z}, \widetilde{T}, \Delta_1, \Delta)$ where $\widetilde{Z} = \min(Z, C)$, $\widetilde{T} = \min(T, C)$, $\Delta_1 = I(Z \leq C)$ and $\Delta = I(T \leq C)$. Here C denotes the potential censoring time, which we assume to be independent of the process (that is, C and (Z, T) are assumed to be independent).

Given two time points $s < t$, define the transition probabilities as $p_{hj}(s,t) = \mathsf{P}(X(t) = j | X(s) = h)$. The transition between the three stochastic states is illustrated in Figure 1. There are five different transition probabilities to estimate: $p_{11}(s,t)$, $p_{12}(s,t)$, $p_{13}(s,t)$, $p_{22}(s,t)$ and $p_{23}(s,t)$. However, only three of them need to be estimated since the two other transition probabilities can be obtained from the following relations: $p_{11}(s,t) + p_{12}(s,t) + p_{13}(s,t) = 1$ and $p_{22}(s,t) + p_{23}(s,t) = 1$.

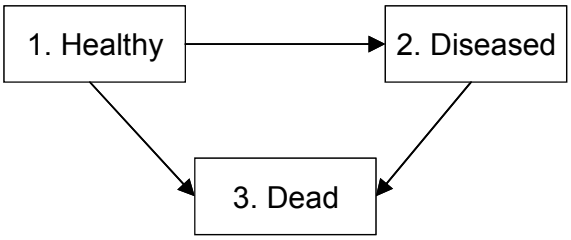

Figure 1: Illness-death model.

In Markov models, the transition probabilities can be calculated from the transition intensities (that we shall assume exist) that we express as

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{p_{hj}(t, t + \Delta t)}{\Delta t}$$

by solving the so-called forward Kolmogorov differential equation (Cox and Miller 1965). For the illness-death model the transition probabilities have explicit expressions,

$$\begin{aligned}
p_{11}(s,t) &= \exp(-A_{12}(s,t) - A_{13}(s,t)), \\
p_{22}(s,t) &= \exp(-A_{23}(s,t)), \\
p_{12}(s,t) &= \int_s^t p_{11}(s,u)\alpha_{12}(u)p_{22}(u,t)\,\mathrm{d}u,
\end{aligned}$$

where $A_{hj}(s,t) = \int_s^t \alpha_{hj}(u)\,\mathrm{d}u$ is the cumulative or integrated intensity between s and t.

In time-homogeneous Markov models the explicit expressions for the transition probabilities are given by

$$\begin{aligned}
p_{11}(s,t) &= \exp(-\alpha_{12}(t-s) - \alpha_{13}(t-s)), \\
p_{22}(s,t) &= \exp(-\alpha_{23}(t-s)), \\
p_{12}(s,t) &= \frac{\alpha_{12}}{\alpha_{12} + \alpha_{13} - \alpha_{23}}[\exp(-\alpha_{23}(t-s)) - \exp(-(\alpha_{12} + \alpha_{13})(t-s))].
\end{aligned}$$

Details about the inference for the transition intensities can be seen in Andersen and Perme (2008).

The transition probabilities can also be estimated nonparametrically or semiparametrically using the notation shown in the top of this section. The expressions for the transition probabilities are given by

$$p_{11}(s,t) = \frac{\mathsf{P}(Z > t)}{\mathsf{P}(Z > s)}, \qquad p_{12}(s,t) = \frac{\mathsf{P}(s < Z \leq t, T > t)}{\mathsf{P}(Z > s)},$$
$$p_{13}(s,t) = \frac{\mathsf{P}(Z > s, T \leq t)}{\mathsf{P}(Z > s)}, \qquad p_{22}(s,t) = \frac{\mathsf{P}(Z \leq s, T > t)}{\mathsf{P}(Z \leq s, T > s)},$$
$$p_{23}(s,t) = \frac{\mathsf{P}(Z \leq s, s < T \leq t)}{\mathsf{P}(Z \leq s, T > s)}.$$

2.1. Aalen-Johansen estimator

The transition probabilities may be estimated via the nonparametric (Aalen-Johansen estimator) model. This can be thought as the generalization of the Kaplan-Meier approach (Kaplan and Meier 1958) for the simple mortality model and was proposed by Aalen and Johansen (1978) for general non-homogeneous Markov models with a finite number of states. Explicit formulae of the Aalen-Johansen estimator for the illness-death model are available (Borgan 1998). Here we give alternative expressions for this estimator using the notation introduced above. The Aalen-Johansen (`AJ`) estimate of the transition probability $p_{11}(s,t)$ is the Kaplan-Meier estimator

$$\hat{p}_{11}^{\mathtt{AJ}}(s,t) = \prod_{s < \widetilde{Z}_i \leq t} \left[1 - \frac{\Delta_{1i}}{n\widetilde{M}_{0n}(\widetilde{Z}_i)}\right], \tag{1}$$

where

$$\widetilde{M}_{0n}(y) = \frac{1}{n}\sum_{j=1}^{n} I(\widetilde{Z}_j \geq y).$$

Similarly, the estimate of the transition probability $p_{22}(s,t)$ is the Kaplan-Meier estimator

$$\hat{p}_{22}^{\mathtt{AJ}}(s,t) = \prod_{s < \widetilde{T}_i \leq t, \widetilde{Z}_i < \widetilde{T}_i} \left[1 - \frac{\Delta_i}{n\widetilde{M}_{1n}(\widetilde{T}_i)}\right], \tag{2}$$

where

$$\widetilde{M}_{1n}(y) = \frac{1}{n}\sum_{j=1}^{n} I(\widetilde{Z}_j < y \leq \widetilde{T}_j).$$

Finally, the estimator for $p_{12}(s,t)$ is given by

$$\hat{p}_{12}^{\mathtt{AJ}}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{p}_{11}^{\mathtt{AJ}}(s,\widetilde{Z}_i^-)\hat{p}_{22}^{\mathtt{AJ}}(\widetilde{Z}_i,t) I(s < \widetilde{Z}_i \leq t, \widetilde{Z}_i < \widetilde{T}_i)}{n\widetilde{M}_{0n}(\widetilde{Z}_i)}, \tag{3}$$

where

$$\hat{p}_{11}^{\mathtt{AJ}}(s,t^-) = \lim_{u \uparrow t} \hat{p}_{11}^{\mathtt{AJ}}(s,u).$$

2.2. Presmoothed Aalen-Johansen estimator

The standard error of the Aalen-Johansen estimator may be large when the censoring is heavy, particularly with a small sample size. Interestingly, the variance of the Aalen-Johansen estimator may be reduced by presmoothing (Dikta 1998). Presmoothing the Aalen-Johansen estimator (Moreira *et al.* 2013) involves replacing the censoring indicators (in the transition probabilities $p_{11}(s,t)$ and $p_{22}(s,t)$) by a smooth fit (e.g., using logistic regression). Then, the corresponding presmoothed Aalen-Johansen (`PAJ`) estimator of $p_{11}(s,t)$ is given by

$$\hat{p}_{11}^{\texttt{PAJ}}(s,t) = \prod_{s<\widetilde{Z}_i\leq t}\left[1-\frac{m_{0n}(\widetilde{Z}_i)}{n\widetilde{M}_{0n}(\widetilde{Z}_i)}\right], \tag{4}$$

where $m_{0n}(\widetilde{Z})$ stands for an estimator of the conditional probability of the event $\Delta_1 = 1$ given $\widetilde{Z}$; which can be estimated using logistic regression. The presmoothed version of (2) given by

$$\hat{p}_{22}^{\texttt{PAJ}}(s,t) = \prod_{s<\widetilde{T}_i\leq t,\widetilde{Z}_i<\widetilde{T}_i}\left[1-\frac{m_{1n}(\widetilde{Z}_i,\widetilde{T}_i)}{n\widetilde{M}_{1n}(\widetilde{T}_i)}\right], \tag{5}$$

where $m_{1n}(\widetilde{Z},\widetilde{T})$ stands for an estimator of the conditional probability of the event $\Delta = 1$ given $(\widetilde{Z},\widetilde{T})$ and given that transition $1 \to 2$ is observed. Finally the transition probability $p_{12}(s,t)$ can be estimated by plugging (4) and (5) into Equation 3.

In the limit case of no presmoothing, the presmoothed Aalen-Johansen estimator reduces to the time-honored Aalen-Johansen estimator, which has become the standard tool for estimating the transition probabilities in Markovian processes. Moreira *et al.* (2013) derive the consistency of the `PAJ` estimator which may be much more efficient than the original `AJ` estimator.

The original and the presmoothed `AJ` estimators are consistent in Markov models. If the Markov property assumption is violated, then the consistency of the time-honored Aalen-Johansen estimator and of its presmoothed version cannot be ensured in general. Alternative methods that do not rely on the Markov assumption are presented below.

2.3. Kaplan-Meier weighted estimator

Recently Meira-Machado *et al.* (2006) verified that in non-Markov situations, the use of Aalen-Johansen estimators to empirically estimate the transition probabilities may be inappropriate. These authors propose, in the scope of the illness-death model, alternative "Markov-free" estimators for the transition probabilities, which do not rely on the Markov assumption. The idea behind estimation is to use the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. The proposed estimator (Kaplan-Meier weighted estimator, `KMW`) is given by

$$\hat{p}_{11}^{\texttt{KMW}}(s,t) = \frac{\sum_{i=1}^{n} W_{1i} I(\widetilde{Z}_i > t)}{\sum_{i=1}^{n} W_{1i} I(\widetilde{Z}_i > s)}, \tag{6}$$

$$\hat{p}_{12}^{\texttt{KMW}}(s,t) = \frac{\sum_{i=1}^{n} W_i I(s < \widetilde{Z}_i \leq t, \widetilde{T}_i > t)}{\sum_{i=1}^{n} W_{1i} I(\widetilde{Z}_i > s)}, \tag{7}$$

$$\hat{p}_{22}^{\texttt{KMW}}(s,t) = \frac{\sum_{i=1}^{n} W_i I(\widetilde{Z}_i \leq s, \widetilde{T}_i > t)}{\sum_{i=1}^{n} W_i I(\widetilde{Z}_i \leq s, \widetilde{T}_i > t)}, \tag{8}$$

where W_i (and W_{1i}) are Kaplan-Meier weights attached to $\widetilde{T}_i$ (respectively, $\widetilde{Z}_i$) when estimating the marginal distribution of T (respectively, Z) from $(\widetilde{T}_i, \Delta_i)$'s (respectively, $(\widetilde{Z}_i, \Delta_{1i})$). The expression for the Kaplan-Meier weights, W_i, is given by $W_i = \frac{\Delta_i}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_j}{n-j+1}\right]$. Meira-Machado *et al.* (2006) derive large sample properties of these estimators which may be generalized to more complicated non-Markov processes.

2.4. Kaplan-Meier presmooth weighted estimator

Recently, Amorim *et al.* (2011) propose a modification of estimator (6)–(8) based on presmoothing, which allows for a variance reduction in the presence of censoring. Basically, this method is obtained by replacing the censoring indicator variables in the expression of the Kaplan-Meier weights by a smooth fit of a binary regression. In this estimator (Kaplan-Meier presmooth weighted estimator, `KMPW`) the presmoothed Kaplan-Meier weights are given by

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

Here, $m(x, y) = \mathsf{P}(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{T} = y, \Delta_1 = 1)$. $m(\widetilde{T}_1, \widetilde{T})$ belongs to a parametric (smooth) family of binary regression curves, e.g., logistic. Our package provides the results assuming that m denotes a logistic regression model (`KMPW`). 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 $(\widetilde{T}_1, \widetilde{T})$ for those with $\Delta_1 = 1$. In the limit case of no presmoothing, the `KMPW` estimator reduces to the `KMW` estimator. Conditions under which both estimators are consistent is fully discussed in papers by Meira-Machado *et al.* (2006) and Amorim *et al.* (2011). In the latter paper the authors compare the performance of the presmoothed (semiparametric) estimator with the purely nonparametric estimator (without presmoothing) and concluded that the presmoothed estimator gains efficiency. The advantages of presmoothing are more clearly seen with an increasing censoring degree and at the distribution's right tail. In general, presmoothing introduces some bias in estimation, while reducing the variance. This bias component is larger when there is some misspecification in the chosen parametric model. Importantly, the validity of a given model for presmoothing can be checked graphically or formally, by applying a goodness-of-fit test. This implies that the risk of introducing a large bias through a misspecified model can be controlled in practice.

2.5. Inverse probability of censoring weighted estimator

To account for the influence of covariates, Meira-Machado *et al.* (2012) introduce estimation methods for the transition probabilities conditionally on current or past measures which we denote by X. The authors provide two competing nonparametric regression estimators for the conditional transition probabilities, $p_{hj}(s, t|X)$, both valid under certain regularity conditions even when the system is non-Markovian. The two estimators use different schemes of inverse censoring probability reweighting to deal with right censoring. In both estimators, local smoothing is done by introducing regression weights that are either based on a local constant (i.e., Nadaraya-Watson) or a local linear regression. To introduce these estimators, we need to introduce first the distribution function of C given X, G_X. Let G_{X_i} denote the conditional distribution function of $C \mid X = X_i$ and let $\widehat{G}_{X_i}$ stand for its estimator. This can be done

using the estimator introduced by Beran (1981),

$$\widehat{G}_x(t) = \prod_{T_i \le t, \Delta_i = 0} \left[1 - \frac{NW_{0i}(x, a_n)}{\sum_{j=1}^{n} I(T_j \ge T_i) NW_{0j}(x, a_n)} \right], \tag{9}$$

with

$$NW_{0i}(x, a_n) = \frac{K_0\left((x - X_i)/a_n\right)}{\sum_{j=1}^{n} K_0\left((x - X_j)/a_n\right)},$$

where $NW_{0i}(x, a_n)$ are the Nadaraya-Watson (`NW`) weights, K_0 is a known probability density function and a_n is a sequence of bandwidths. This estimator reduces to the so-known Kaplan-Meier (Kaplan and Meier 1958) estimator when all weights are equal. Then, the inverse probability censoring weighted estimators (`IPCW`) are given by

$$\hat{p}_{11}^{\texttt{IPCW}}(s,t|X=x) = \frac{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i > t)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}^-)}}{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i > s)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}^-)}}, \tag{10}$$

$$\hat{p}_{12}^{\texttt{IPCW}}(s,t|X=x) = \frac{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(s < \widetilde{Z}_i \le t, \widetilde{T}_i > t)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}^-)}}{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i > s)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}^-)}}, \tag{11}$$

$$\hat{p}_{22}^{\texttt{IPCW}}(s,t|X=x) = \frac{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i \le s, \widetilde{T}_i > t)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}^-)}}{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i \le s, \widetilde{T}_i > s)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}^-)}}, \tag{12}$$

where $NW_{1i}(x, b_n)$ are `NW` weights as introduced above and $\hat{G}_{X_i}(\widetilde{T}^-) = \hat{G}_{x=X_i}(\widetilde{T}^-)$. Alternatively local linear weights can also be introduced.

An alternative approach that also accounts for the influence of covariates is based on the Lin, Sun, and Ying (1999) approach for the bivariate distribution function. Then, a different set of estimators (`LIN`) are given by

$$\hat{p}_{11}^{\texttt{LIN}}(s,t|X=x) = \frac{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i > t)}{1-\hat{H}_{X_i}(t^-)}}{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i > s)}{1-\hat{H}_{X_i}(s^-)}}, \tag{13}$$

$$\hat{p}_{12}^{\texttt{LIN}}(s,t|X=x) = \frac{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(s < \widetilde{Z}_i \le t, \widetilde{T}_i > t)}{1-\hat{G}_{X_i}(t^-)}}{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i > s)}{1-\hat{G}_{X_i}(s^-)}}, \tag{14}$$

$$\hat{p}_{22}^{\texttt{LIN}}(s,t|X=x) = \frac{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i \le s, \widetilde{T}_i > t)}{1-\hat{G}_{X_i}(t^-)}}{\sum_{i=1}^{n} NW_{1i}(x,b_n) \frac{I(\widetilde{Z}_i \le s, \widetilde{T}_i > s)}{1-\hat{G}_{X_i}(s^-)}}, \tag{15}$$

where $\hat{H}_X$ stands for the Kaplan-Meier estimator of the conditional distribution of C given X based on the $(\widetilde{Z}_i, 1 - \Delta_{1i})$'s. This estimator is defined in the same way as $\hat{G}_x$.

Here we assume that C is independent of (Z, T) given X; this assumption does not exclude the possibility of dependent censoring. The performance of the two estimators has been

evaluated through simulations, showing that they are valid even when the system is non-Markov or conditionally non-Markov. Simulation results show that the general performance difference between the two methods is quite small, and both methods perform quite well. However, one of the two approaches (the LIN-based one) has the drawback of occasionally providing nonmonotone curves for transition probabilities which are indeed monotone and, therefore, its practical use is less recommendable.

2.6. Location-scale estimator

Other estimators were proposed to estimate the transition probabilities. A valid estimator was provided by Keilegom *et al.* (2011). This methodology assumes that the vector of gap times (Z, Y), where $Y = T - Z$, satisfies the nonparametric location-scale regression model, allowing for the transfer of tail information from lightly censored areas to heavily ones. An automatic bandwidth procedure was proposed by Meira-Machado, Roca-Pardiñas, Keilegom, and Cadarso-Suárez (2013) for this methodology.

Consider the nonparametric location-scale regression model (LS) $Y = m(Z) + \sigma(Z)\epsilon$, where the functions m and σ are 'smooth', and ϵ is independent of Z. Under this model, the authors propose a nonparametric estimator of the distribution of the error variable, F_ϵ, to introduce nonparametric estimators for the transition probabilities. They use a Kaplan-Meier estimator of F_ϵ based on the $(\hat{E}_i, \Delta_i)$'s (where $\hat{E}_i = (\widetilde{Y}_i - \hat{m}(\widetilde{Z}i))/\hat{\sigma}(\widetilde{Z}_i)$) which is the key for the construction of an estimator for the conditional distribution of the second gap time, $\hat{F}(y|x) = \hat{F}_\epsilon(\frac{y-\hat{m}(x)}{\hat{\sigma}(x)})$. The location and scale functionals are estimated using an extension of the Beran (1981) estimator, which copes with censoring in the first gap time. Then, estimators for the transition probabilities can be obtained from the following expressions:

$$\begin{aligned}
\hat{p}_{11}^{\mathtt{LS}}(s,t) &= \big(1-\hat{F}_1(t)\big)/\big(1-\hat{F}_1(s)\big),\\
\hat{p}_{12}^{\mathtt{LS}}(s,t) &= \frac{1}{1-\hat{F}_1(s)}\int_s^t \big[1-\hat{F}(t-u|u)\big]\,\hat{F}_1(\mathrm{d}u),\\
\hat{p}_{22}^{\mathtt{LS}}(s,t) &= \frac{\int_0^s \big[1-\hat{F}(t-u|u)\big]\,\hat{F}_1(\mathrm{d}u)}{\int_0^s \big[1-\hat{F}(s-u|u)\big]\,\hat{F}_1(\mathrm{d}u)},
\end{aligned}$$

where $F_1(\cdot)$ is the marginal distribution of the first gap time, which we may estimate by the Kaplan-Meier estimator based on the $(\tilde{Z}_i, \Delta_{1i})$'s.

Simulations reported in Meira-Machado *et al.* (2013) suggest that the transfer of tail information may improve the estimation of the transition probabilities especially in points where the uncensored information is scarce. The authors compared the location-scale method with the estimator by Meira-Machado *et al.* (2006) in several scenarios. It was found that when the deviation from the location-scale model was only minor, the location-scale method outperforms the Kaplan-Meier weighted estimator (Meira-Machado *et al.* 2006). However, when the model deviates a lot from a location-scale model, the later method becomes better. Another drawback of the location-scale model is that this method can only be used in the progressive three-state model.

2.7. Occupation probabilities

Another important target in multi-state modeling is the estimation of the state occupation probabilities. For the illness-death model there are in essence three state occupation proba-

bilities to calculate, $p_{11}(0,t)$, $p_{12}(0,t)$ and $p_{13}(0,t)$. Datta and Satten (2001) show that these quantities can be estimated using Aalen-Johansen estimators even when the process is not Markov. Though all methods introduced in the previous sections provide valid estimators for these quantities, the Markovian approaches (`AJ` and `PAJ`) are recommended.

3. Package description

The **TPmsm** software package contains functions that calculate estimates for the transition probabilities. As mentioned in Section 2, this software package can be used to implement seven methods (`AJ`, `PAJ`, `KMW`, `KMPW`, `IPCW`, `LIN` and `LS`). This software package is intended to be used within the statistical software program R (R Core Team 2014). **TPmsm** is composed of several functions that allow users to obtain estimates and plots of the transition probabilities. Table 1 provides a summary of some of the functions in the package. Details on the usage of these functions can be obtained with the corresponding help pages.

Function	Description
`dgpTP`	Generate data from an illness-death model (based on some known copula functions). By default returns a data set of class '`survTP`'.
`corrTP`	Correlation between the bivariate times for some copula distributions.
`survTP`	Set up adequate data set of class '`survTP`' for implementing all the methods.
`transAJ`	Aalen-Johansen (`AJ`) estimates for the transition probabilities.
`transPAJ`	Presmoothed Aalen-Johansen (`PAJ`) estimates for the transition probabilities.
`transKMW`	Kaplan-Meier weighted (`KMW`) estimates for the transition probabilities.
`transKMPW`	Kaplan-Meier presmoothed weighted (`KMPW`) estimates for the transition probabilities.
`transIPCW`	Inverse probability of censoring weighted (`IPCW`) estimates for the transition probabilities.
`transLIN`	LIN-based (`LIN`) estimates for the transition probabilities.
`transLS`	Location-scale (`LS`) estimates for the transition probabilities.
`plot`	Plots for the transition probabilities.
`setThreadsTP`	Specifies the number of threads used by default in parallel sections.
`setPackageSeedTP`	Set the initial package seed.

Table 1: Summary of functions in the package.

It should be noted that to apply the methods described in Section 2 one needs the following variables: `time1`, `event1`, `Stime` and `event`. A single covariate can also be included (they are necessary only for `IPCW` and `LIN` methods). The variable `time1` represents the observed time in State 1 ("healthy"), and `event1` the corresponding status/censoring indicator (if the survival time is a censored observation, the value is 0 and otherwise the value is 1). The variable `Stime` represents the total survival time (time to the absorbing state). If `event1 = 0`, then the total survival time is equal to the observed time in State 1. The variable `event` is the final status of the individual (takes the value 1 if the final event of interest is observed and 0 otherwise).

4. Data generation

Users may use the function `dgpTP` to generate data from the illness-death model. We assume that all individuals are in the "healthy" state at time $t = 0$. Therefore, the patient's history (or course) may be divided into two groups according to whether the disease occurred (that is, passing through State 2) $(1 \to 2 \to 3)$ or not $(1 \to 3)$. We separately consider these two possible subgroups of individuals. For the first subgroup of individuals, the successive gap times $(Z, T - Z)$ can be simulated from two of the most known copula functions: the Farlie-Gumbel-Morgenstern copula with exponential marginals and the bivariate Weibull distribution.

In the following, using the `dgpTP` function we will simulate data from the illness-death model using Gumbel's bivariate exponential distribution (`dist = "exponential"`) $F_{12}(x, y) = F_1(x)F_2(y)\left[1 + \theta\left\{1 - F_1(x)\right\}\left\{1 - F_2(y)\right\}\right]$ with unit exponential margins (`dist.par = c(1, 1)`). The parameter θ controls for the amount of dependency between the gap times $(Z, T - Z)$. Theoretical correlation between the gap times can be obtained using the `corrTP` function. For the second subgroup of individuals (those that go directly from State 1 to State 3), the corresponding survival time is simulated according to an exponential with rate parameter 1.

The computation and the implementation of the proposed estimator involves the construction of pointwise confidence intervals by means of a bootstrap approach and in some cases the choice of an appropriate bandwidth. Thus, some of the methods implemented in package **TPmsm** can be computationally demanding. To obtain the point estimation and the pointwise confidence intervals, efficient algorithms were developed and implemented in the C programming language. The most computationally demanding parts of the code, namely those that involve the bootstrap and cross-validation techniques, were parallelized by means of the OpenMP API. This should considerably increase performance on multi-core/multi-threading machines. To ensure the reproducibility of the results reported in the paper, two threads were considered (`setThreadsTP(2)`). The random number generator with multiple independent streams ((L'Ecuyer 1999),(L'Ecuyer, Simard, Chen, and Kelton 2002) and (Karl, Eubank, Milovanovic, Reiser, and Young 2014)) was implemented for parallel computation of uniform pseudorandom numbers. Package **TPmsm** own implemementation of a random number generator makes it independent of R, requiring a different function for defining a seed. The function `setPackageSeedTP` requires a vector of six integers.

```
library("TPmsm");
setThreadsTP(2);
seed <- c(2718, 3141, 5436, 6282, 8154, 9423);
setPackageSeedTP(seed);
sim_data_exp <- dgpTP(n = 1000, corr = 0, dist = "exponential",
  dist.par = c(1, 1), model.cens = "uniform", cens.par = 3,
  state2.prob = 0.5);
```

This input command will simulate 1000 observations ($n = 1000$) assuming no correlation in Gumbel's bivariate exponential distribution (`corr = 0`), using an independent uniform censoring time (`model.cens = "uniform"`), according to model $U(0, 3)$ (`cens.par = 3`). The use of `corr = 0` in Gumbel's bivariate exponential distribution leads to independent gap times and therefore to Markov data. The proportion of transitions into State 2 is given by the argument `state2.prob` (a value of 1 corresponds to the progressive three-state model).

To obtain the estimates for the methods proposed in Section 2 we can use the functions shown in Table 1. As in the simulation by Amorim *et al.* (2011) and Moreira *et al.* (2013) we are going to obtain estimates for transition probabilities at values `s = 0.5108` and `t = 0.9163`. The true values for the transition probabilities are: $p_{11}(s,t) = \frac{\mathsf{P}(Z>t)}{\mathsf{P}(Z>s)} = 0.667$, $p_{12}(s,t) = \frac{\mathsf{P}(s<Z\leq t, T>t)}{\mathsf{P}(Z>s)} = 0.135$ and $p_{22}(s,t) = \frac{\mathsf{P}(Z\leq s, T>t)}{\mathsf{P}(Z\leq s, T>s)} = 0.666$. The following two input commands provide the estimates for the `AJ` and `PAJ` methods. Since the process is Markovian these are the recommended approaches. With these input commands we obtain the estimates for the transition matrix together with 95% (`conf.level = 0.95`) pointwise confidence intervals (`conf = TRUE`) using 1000 bootstrap replicates (`n.boot = 1000`). The construction of the pointwise confidence intervals is obtained by randomly sampling the n items from the original data set with replacement. This can be achieved using the percentile bootstrap interval (`method.boot = "percentile"`) or using the basic bootstrap interval (`method.boot = "basic"`). By default all functions use the percentile bootstrap method (Davison and Hinkley 1997).

```
transAJ(object = sim_data_exp, s = 0.5108, t = 0.9163, conf = TRUE,
  conf.level = 0.95, n.boot = 1000);

## Aalen-Johansen transition probabilities
##
## Estimates of P(0.5108, 0.9163)
##           1         2         3
## 1 0.6479348 0.1813716 0.1706936
## 2 0.0000000 0.7539452 0.2460548
## 3 0.0000000 0.0000000 1.0000000
##
## Bootstrap confidence bands with 1000 samples
##
## 2.5%
##           1         2         3
## 1 0.6055291 0.1494781 0.1401043
## 2 0.0000000 0.6854783 0.1852007
## 3 0.0000000 0.0000000 1.0000000
##
## 97.5%
##           1         2         3
## 1 0.6929826 0.2135551 0.2029478
## 2 0.0000000 0.8147993 0.3145217
## 3 0.0000000 0.0000000 1.0000000

transPAJ(object = sim_data_exp, s = 0.5108, t = 0.9163, conf = TRUE,
  conf.level = 0.95, n.boot = 1000);

## Presmoothed Aalen-Johansen transition probabilities
##
## Estimates of P(0.5108, 0.9163)
```

```
##           1         2         3
## 1 0.6718558 0.1841269 0.1440172
## 2 0.0000000 0.7314865 0.2685135
## 3 0.0000000 0.0000000 1.0000000
##
## Bootstrap confidence bands with 1000 samples
##
## 2.5%
##           1         2         3
## 1 0.6350731 0.1493700 0.1132212
## 2 0.0000000 0.6702976 0.2105773
## 3 0.0000000 0.0000000 1.0000000
##
## 97.5%
##           1         2         3
## 1 0.7105014 0.2179436 0.1747968
## 2 0.0000000 0.7894227 0.3297024
## 3 0.0000000 0.0000000 1.0000000
```

Results reveal accuracy for both methods for which the true values are within the bootstrap confidence bands. The bootstrap confidence bands are narrower in the case of the presmoothed Aalen-Johansen estimator revealing less variability for this method. In general, the results for the lower and upper bounds of the bootstrap confidence interval greatly depend on the sample size of the data set and the number of bootstrap simulations. In this case, a second and a third set of 1000 resamples gave similar results for the bootstrap confidence intervals, suggesting that the number of resamples are enough. The CPU time needed for running the `transAJ` function varies depending on whether bootstrap confidence bands are requested or not, the sample size, and the type of processor in the computer. The command presented above took no more than 1 second on a PC with a four Core Intel i7 processor with 8 GB memory. The same input command but with $n = 10000$ resamples took less than a few seconds.

Non-Markov data can also be generated using correlated gap times in Gumbel's bivariate exponential distribution. For example, using a maximum correlation of 25% (using `corr = 1` in the `dgpTP` function) as shown below.

```
setPackageSeedTP(seed);
sim_data_exp2 <- dgpTP(n = 1000, corr = 1, dist = "exponential",
  dist.par = c(1, 1), model.cens = "uniform", cens.par = 3,
  state2.prob = 0.5);
```

The following input commands provide the estimates (with bootstrap confidence bands) for the `KMW` and `KMPW` methods at values `s = 0.5108` and `t = 0.9163`. The true values for the transition probabilities at these values are: $p_{11}(s,t) = 0.667$, $p_{12}(s,t) = 0.134$ and $p_{22}(s,t) = 0.558$. Since the process is not Markov these are the recommended approaches.

```
transKMW(object = sim_data_exp2, s = 0.5108, t = 0.9163, conf = TRUE,
  conf.level = 0.95, n.boot = 1000);

## Kaplan-Meier Weighted transition probabilities
##
## Estimates of P(0.5108, 0.9163)
##           1         2         3
## 1 0.6479348 0.1837512 0.1683140
## 2 0.0000000 0.5140501 0.4859499
## 3 0.0000000 0.0000000 1.0000000
##
## Bootstrap confidence bands with 1000 samples
##
## 2.5%
##           1         2         3
## 1 0.6055291 0.1504573 0.1340369
## 2 0.0000000 0.4039540 0.3734877
## 3 0.0000000 0.0000000 1.0000000
##
## 97.5%
##           1         2         3
## 1 0.6929826 0.2194789 0.2042371
## 2 0.0000000 0.6265123 0.5960460
## 3 0.0000000 0.0000000 1.0000000

transKMPW(object = sim_data_exp2, s = 0.5108, t = 0.9163, conf = TRUE,
  conf.level = 0.95, n.boot = 1000);

## Presmoothed Kaplan-Meier Weighted transition probabilities
##
## Estimates of P(0.5108, 0.9163)
##           1         2         3
## 1 0.6718558 0.1630406 0.1651035
## 2 0.0000000 0.5337576 0.4662424
## 3 0.0000000 0.0000000 1.0000000
##
## Bootstrap confidence bands with 1000 samples
##
## 2.5%
##           1         2         3
## 1 0.6350731 0.1358723 0.1377945
## 2 0.0000000 0.4357253 0.3729777
## 3 0.0000000 0.0000000 1.0000000
##
## 97.5%
##           1         2         3
```

```
## 1 0.7105014 0.1907409 0.1914515
## 2 0.0000000 0.6270223 0.5642747
## 3 0.0000000 0.0000000 1.0000000
```

Results reveal that both methods perform very well. As expected, the presmooth method achieved less variability, with narrower bootstrap confidence bands. Results for the Aalen-Johansen type estimators (`AJ` and `PAJ`) reveal a systematic bias for the transition from State 2 to State 3 (results not shown).

In addition to the numerical results graphical output can also be obtained. This will be shown in the next section using two data sets: the widely used and well-known colon cancer data and data from a bladder cancer study. Details about these data sets are given below.

5. Examples of application

To illustrate our estimators we consider two real data sets. One of these data sets comes from the well-known colon cancer study which is freely available as part of the R **survival** package (Therneau and Grambsch 2000; Therneau 2014). In addition to this data set we also use data from a bladder cancer study (Byar 1980) conducted by the Veterans Administration Cooperative Urological Research Group.

5.1. Colon cancer data

For illustration, we apply some of the proposed methods of Section 2 to data from a large clinical trial on Duke's stage III patients, affected by colon cancer, that underwent a curative surgery for colorectal cancer (Moertel, Fleming, McDonald *et al.* 1990). In this study, some of these patients have residual cancer, which leads to disease recurrence and death (in some cases). From the total of 929 patients, 468 developed a recurrence and among these 414 died. 38 patients have died of causes unrelated to their disease and without evidence of recurrence. The remaining 423 patients contributed with censored survival times. For each individual, an indicator of his/her final vital status (censored or not), the survival times (time to recurrence, time to death) from the entry of the patient in the study (in days), and a vector of covariates including *age* (in years) and *recurrence* (coded as 1 = yes; 0 = no) were recorded. The covariate *recurrence* is a time-dependent covariate which can be expressed as an intermediate event and modeled using the illness-death model with states "alive and disease-free", "alive with recurrence" and "dead".

By including covariates depending on the history (using a Cox proportional hazards model), we verified that the mortality transition for recurring patients is affected by the time spent in the previous state (p value < 0.001). This allowed us to conclude that the Markov assumption may be unsatisfactory for the colon cancer data set and that, consequently, Aalen-Johansen type estimators should not be used. Thus, in this section we illustrate the use of two "Markov-free" estimators (`KMW` and `KMPW`) as well as two additional estimators (`IPCW` and `LIN`) that were proposed to estimate the transition probabilities conditionally on current or past covariate measures such as *age*.

Below is an excerpt of the data with one row per individual.

```
data("colonTP", package = "TPmsm");
head( head(colonTP[ , c(1:4, 7)]) );

##   time1 event1 Stime event age
## 1   968      1  1521     1  43
## 2  3087      0  3087     0  63
## 3   542      1   963     1  71
## 4   245      1   293     1  66
## 5   523      1   659     1  69
## 6   904      1  1767     1  57
```

Each line represents the information from one individual in the study. The variable `time1` denotes the sojourn time in State 1 whereas `Stime` is the total time of survival. `event1` and `event` are the corresponding indicator statuses. Among the first six individuals, only individual represented by line 2 remains alive (and without having had a recurrence) at the end of the study. All the remaining individuals had a recurrence and died before the end of the study. For example, the individual represented by line 1 had a recurrence at day 968 and died at day 1521. Note that `time1` < `Stime` means that a transition from State 1 to State 2 (i.e., recurrence) occurred.

We computed the estimated values for the transition probabilities $p_{hj}(s,t)$ for several pairs (s,t), $s < t$. For illustration purposes we only report the estimated values of $p_{hj}(365, 1096)$ (one year and three years) for the `KMW` and `KMPW` methods with 95% bootstrap confidence intervals.

```
colon_obj <- with( colonTP, survTP(time1, event1, Stime, event, age) );
colon_obj_TP <- transKMW(object = colon_obj, s = 365, t = 1096,
  conf = TRUE, conf.level = 0.95);
colon_obj_TP;

## Kaplan-Meier Weighted transition probabilities
##
## Estimates of P(365, 1096)
##           1         2         3
## 1 0.7192603 0.1432380 0.1375017
## 2 0.0000000 0.1570985 0.8429015
## 3 0.0000000 0.0000000 1.0000000
##
## Bootstrap confidence bands with 1000 samples
##
## 2.5%
##           1         2         3
## 1 0.6853757 0.1177304 0.1114069
## 2 0.0000000 0.1020286 0.7848585
## 3 0.0000000 0.0000000 1.0000000
##
## 97.5%
```

```
##           1         2         3
## 1 0.752907 0.1693951 0.1646936
## 2 0.000000 0.2151415 0.8979714
## 3 0.000000 0.0000000 1.0000000

colon_obj2_TP <- transKMPW(object = colon_obj, s = 365, t = 1096,
  conf = TRUE, conf.level = 0.95);
colon_obj2_TP;

## Presmoothed Kaplan-Meier Weighted transition probabilities
##
## Estimates of P(365, 1096)
##           1         2         3
## 1 0.7194552 0.1433486 0.1371961
## 2 0.0000000 0.1582020 0.8417980
## 3 0.0000000 0.0000000 1.0000000
##
## Bootstrap confidence bands with 1000 samples
##
## 2.5%
##           1         2         3
## 1 0.6853516 0.1182992 0.1104852
## 2 0.0000000 0.1050996 0.7818987
## 3 0.0000000 0.0000000 1.0000000
##
## 97.5%
##           1         2         3
## 1 0.7511722 0.1699785 0.1598693
## 2 0.0000000 0.2181013 0.8949004
## 3 0.0000000 0.0000000 1.0000000
```

The outputs for the transition probabilities could be useful in understanding the patients' illness stage over time. Plots for these quantities can easily be obtained. Figure 2 plots the transition probabilities $p_{hj}(365, t)$ for all allowed transitions using the `KMW` method. This plot can be obtained using the following input commands:

```
colon_obj_TP <- transKMW(object = colon_obj, s = 365, conf = TRUE,
  conf.level = 0.95);
plot(colon_obj_TP, col = seq_len(5), lty = 1, ylab = "p_hj(365,t)");
```

Figure 3 depicts the `KMW` estimates of $p_{12}(s = 365, t)$ as functions of the time (for a fixed value of $s = 365$) together with a 95% pointwise confidence band based on simple bootstrap. The estimates shown in the main curve indicate that this probability increases until around time $t = 600$ and afterwards decreases.

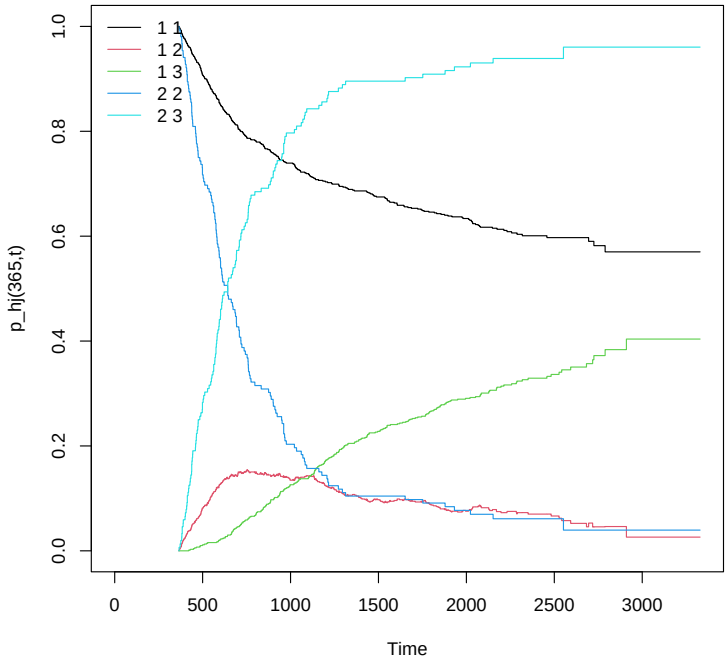

Figure 2: Transition probability estimates using the `KMW` method. Colon cancer data.

```
plot(colon_obj_TP,  tr.choice = "1 2", conf.int = TRUE, ylim = c(0, 0.2),
  legend = FALSE, ylab = "p12(365,t)");
```

Estimates for the conditional transition probabilities can be obtained using two methods (`IPCW` and `LIN`). Below we present the input command to obtain the estimates for the `IPCW` method for a vector of two `ages` (40 and 68). Results suggest a real influence of the covariate `age` in the survival prognosis. More specifically, patients with 40 years have a larger probability of recurrence than patients with 68 years. Note that the estimate obtained for those patients with 40 years is not within the bootstrap confidence bands obtained for those with 68 years. These insights can also be seen in Figures 4 and 5 which depict respectively the `IPCW` estimates of the conditional transition probabilities $p_{11}(365, 1096|\texttt{age})$ and $p_{12}(365, 1096|\texttt{age})$ as functions of the covariate `age` together with a 95% pointwise confidence band based on simple bootstrap. In both plots it is seen that these curves are not constant. Furthermore, the effects of `age` depicted in Figure 5, suggest a real influence of age on survival. More specifically, patients near the forties have a larger probability of recurrence than older patients. Note that it would not be possible to include a horizontal line within the confidence bands in this plot. An alternative method that accounts for the influence of continuous covariates is the `LIN` method which is implemented in the `transLIN` function. Similarly, `transIPCW` can also handle one covariate.

```
CTP_obj <- transIPCW(colon_obj, s = 365, t = 1096, x = c(40, 68),
  conf = TRUE, n.boot = 1000, method.boot = "percentile");
CTP_obj;
```

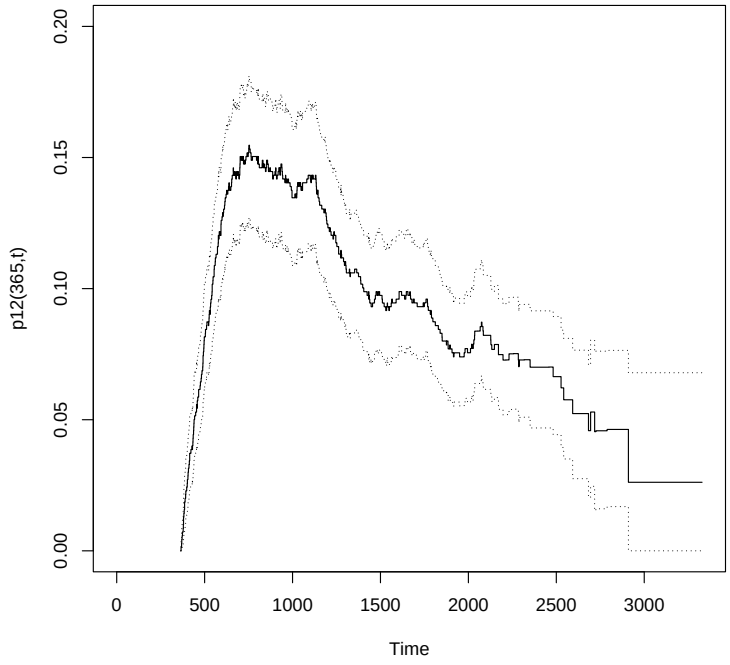

Figure 3: Transition probability estimates, with bootstrap confidence bands, using the `KMW` method. Colon cancer data.

```
## Inverse Probability Censoring Weighted conditional transition probabilities
##
## Estimates of P(365, 1096 | 40)
##           1         2         3
## 1 0.6586309 0.3004002 0.0409689
## 2 0.0000000 0.2063084 0.7936916
## 3 0.0000000 0.0000000 1.0000000
##
## Bootstrap confidence bands with 1000 samples
##
## 2.5%
##           1         2           3
## 1 0.5358529 0.1925062 0.007045676
## 2 0.0000000 0.0000000 0.561993075
## 3 0.0000000 0.0000000 1.000000000
##
## 97.5%
##           1         2          3
## 1 0.7732796 0.4232946 0.08588793
## 2 0.0000000 0.4380069 1.00000000
## 3 0.0000000 0.0000000 1.00000000
```

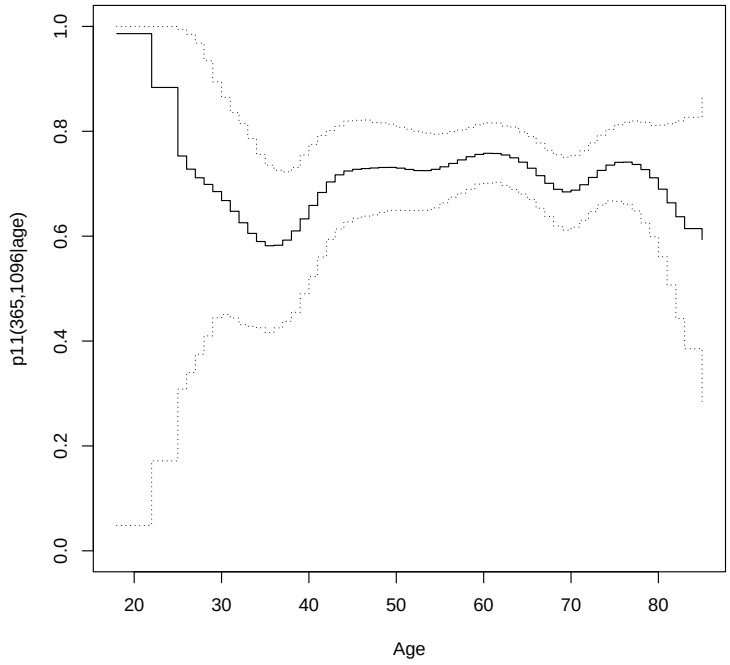

Figure 4: Evolution of the transition probability $p_{11}(365, 1096)$ along the covariate `age` with 95% bootstrap confidence bands based on the `IPCW` method. Colon cancer data.

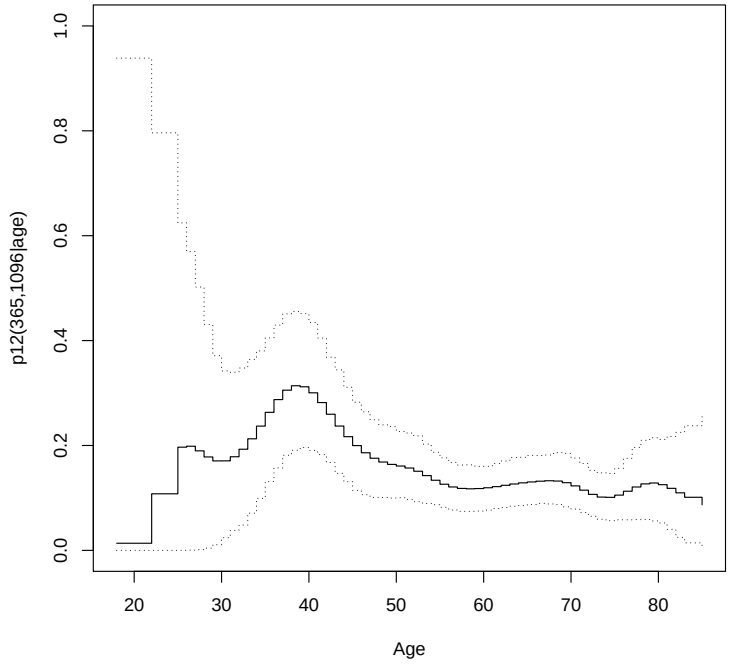

Figure 5: Evolution of the transition probability $p_{12}(365, 1096)$ along the covariate `age` with 95% bootstrap confidence bands based on the `IPCW` method. Colon cancer data.

```
##
## Estimates of P(365, 1096 | 68)
##           1         2         3
## 1 0.6893005 0.1321605 0.1785390
## 2 0.0000000 0.1325104 0.8674896
## 3 0.0000000 0.0000000 1.0000000
##
## Bootstrap confidence bands with 1000 samples
##
## 2.5%
##          1          2         3
## 1 0.618202 0.08461320 0.1259819
## 2 0.000000 0.04160551 0.7545302
## 3 0.000000 0.00000000 1.0000000
##
## 97.5%
##           1         2         3
## 1 0.7500113 0.1845222 0.2377143
## 2 0.0000000 0.2454698 0.9583945
## 3 0.0000000 0.0000000 1.0000000
```

```
plot(CTP_obj, plot.type = "c", tr.choice = "1 1", conf.int = TRUE,
  xlab = "Age", legend = FALSE, ylab = "p11(365,1096|age)");
plot(CTP_obj, plot.type = "c", tr.choice = "1 2", conf.int = TRUE,
  xlab = "Age", legend = FALSE, ylab = "p12(365,1096|age)");
```

Alternatively, we can view all transitions in the same plot using the following input command (Figure 6):

```
plot(CTP_obj, plot.type = "c", col = seq_len(5), lty = 1, xlab = "Age",
  ylab = "p_hj(365,1096|age)");
```

A contour plot of the transition probabilities can be obtained using the `contour` function; a grid of colored or gray-scale rectangles with colors corresponding to the values of the transition probabilities can be obtained using the `image` function. Details on the usage of these functions can be obtained within the corresponding help pages.

5.2. Example of application: Bladder cancer study

The methods described in Section 2.6 are illustrated using data from a bladder cancer study (Byar 1980) conducted by the Veterans Administration Cooperative Urological Research Group. 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

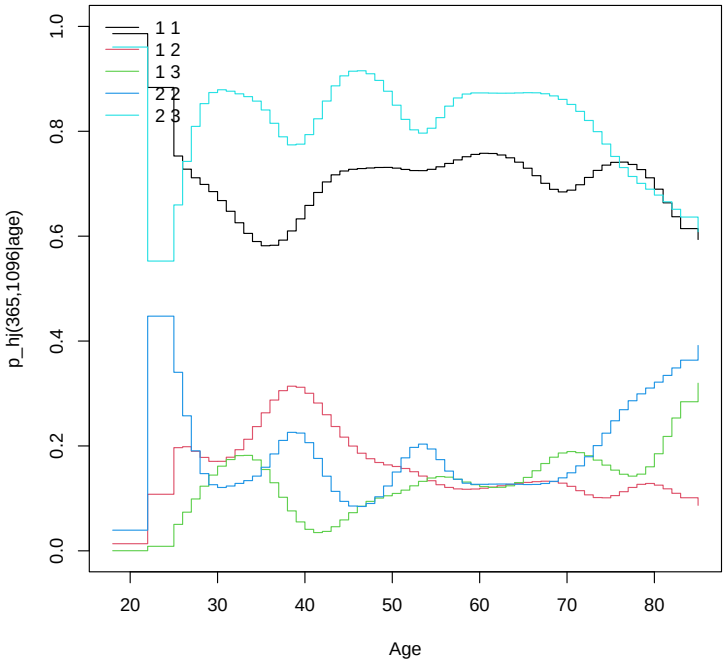

Figure 6: Evolution of the transition probabilities $p_{hj}(365, 1096)$ along the covariate `age`, based on the `IPCW` method. Colon cancer data.

(in months) and the corresponding gap times, Z and $Y = T - Z$, are considered. Thus, we have a progressive three-state model with state "alive and disease-free", "first recurrence" and "second recurrence". From the total of 85 patients, 47 relapsed at least once and, among these, 29 experienced a new recurrence.

For large values of s and t, the transition probabilities $p_{hj}(s, t)$ will be difficult to estimate in a completely nonparametric way. This will be particularly true in situations where censoring percentages are high as for our data set for which we have a total amount of censoring of 66%. The location-scale method is appropriate for the bladder cancer data since this methodology is mainly relevant for estimation in the right tail of the distribution where the censoring effects are strong at those points (uncensored information is scarce).

We will calculate estimates for the transition probabilities in several points and plot these estimates. This will be done using the function `transLS`.

```
data("bladderTP", package = "TPmsm");
head(bladderTP);

##   time1 event1 Stime event
## 1     1      0     1     0
## 2     4      0     4     0
## 3     7      0     7     0
## 4    10      0    10     0
```

```
## 5     6      1    10     0
## 6    14      0    14     0
```

We computed the estimated values for the transition probabilities $p_{hj}(s,t)$ for several pairs (s,t), $s<t$. For illustration purposes we only report the estimated values of $p_{hj}(3,8)$ for the `LS` method with 95% bootstrap confidence intervals. The success of the `LS` method greatly depends on the choice of an appropriate bandwidth. The selection of the optimal bandwidth is highly computationally intensive, but is crucial to the success of the location-scale method. To select the bandwidth we use a weighted cross-validation error criterion, with weights based on the Kaplan-Meier estimator. Details about these procedures can be seen in the paper by Meira-Machado *et al.* (2013). Results for the transition probabilities $p_{hj}(3,8)$ shown below were obtained using a grid of 100 bandwidth values (`nh = 100`) over the interval between 0.0001 and 1 (`h = c(0.0001, 1)`) and considering 100 cross-validation samples (`ncv = 100`).

```
bladderTP_obj <- with( bladderTP, survTP(time1, event1, Stime, event) );
LS_obj <- transLS(object = bladderTP_obj, s = 3, t = 8, h = c(0.0001, 1),
  nh = 100, ncv = 100, conf = TRUE);
LS_obj;

## Location-Scale transition probabilities
##
## Estimates of P(3, 8)
##           1         2           3
## 1 0.8391534 0.1552910 0.005555556
## 2 0.0000000 0.9222722 0.077727794
## 3 0.0000000 0.0000000 1.000000000
##
## Bootstrap confidence bands with 1000 samples
##
## 2.5%
##           1          2           3
## 1 0.7384387 0.06143224 0.000000000
## 2 0.0000000 0.86275399 0.005852838
## 3 0.0000000 0.00000000 1.000000000
##
## 97.5%
##           1         2         3
## 1 0.9310463 0.2554292 0.0467999
## 2 0.0000000 0.9941472 0.1372460
## 3 0.0000000 0.0000000 1.0000000
```

Plots for the transition probabilities can also be obtained. Figure 7 plots the transition probabilities $p_{hj}(3,t)$ for all allowed transitions. In Figure 8 we can see the plot for the transition probability $p_{12}(3,t)$ along the pointwise confidence bands using the `LS` method. These plots are obtained using the following input commands:

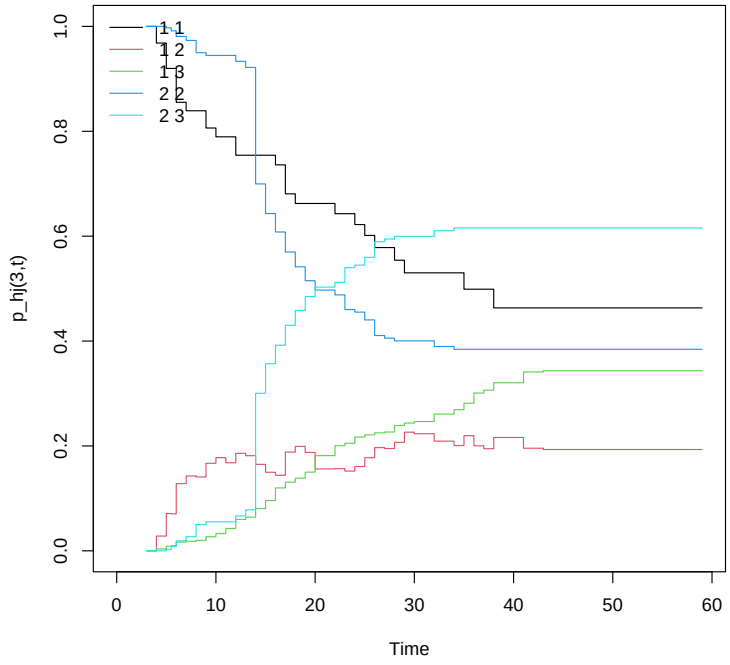

Figure 7: Transition probability estimates using the `LS` method. Bladder cancer data.

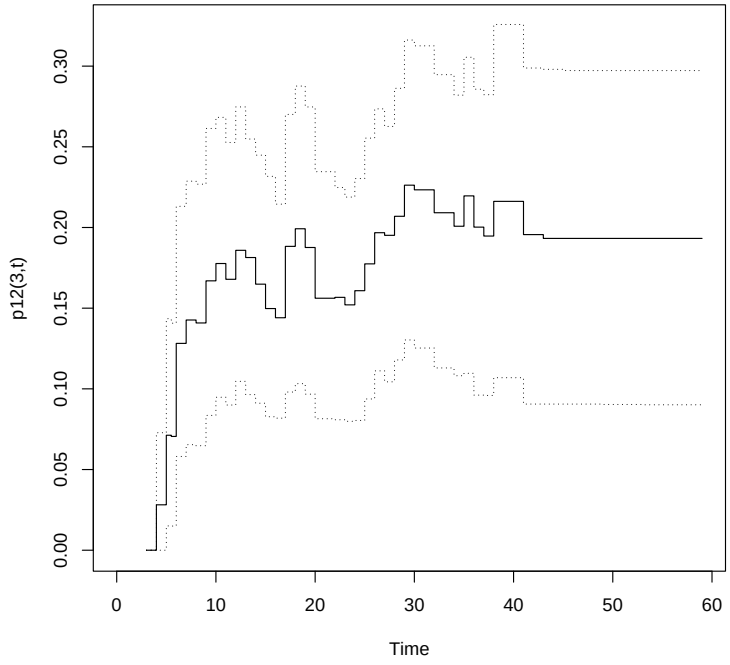

Figure 8: Transition probability estimates, with bootstrap confidence bands, using the `LS` method. Bladder cancer data.

```
LS2_obj <- transLS(object = bladderTP_obj, s = 3, t = 60, h = c(0.0001, 1),
  nh = 100, ncv = 100, conf = TRUE);
plot(LS2_obj, col = seq_len(5), lty = 1, ylab = "p_hj(3,t)");
plot(LS2_obj, tr.choice = "1 2", conf.int = TRUE, ylab = "p12(3,t)",
  ylim = c(0, 0.325), legend = FALSE);
```

6. Conclusion

This paper discusses the implementation of some newly developed methods for the transition probabilities in the illness-death model in an R package. The **TPmsm** package uses seven nonparametric and semiparametric estimators. One of these estimators is the Aalen-Johansen estimator (Aalen and Johansen 1978) under the assumption of a Markovian data generating process. A modification of the Aalen-Johansen estimator (Moreira *et al.* 2013), based on a preliminary estimation (presmoothing) of the censoring probability for the total time, given the available information is also implemented. This method allows for a variance reduction in the presence of censoring, in particular for situations with high percentages of censored total time among the uncensored subjects in State 1.

If there is no evidence against the Markov condition then the time honored Aalen-Johansen estimator and its presmoothed version will be preferred. If the Markov property is violated, then the consistency of these estimators cannot be ensured in general. Exceptions to this are the estimator for the occupation probabilities. Alternative estimators of the transition probabilities not relying on the Markov condition were recently proposed (Meira-Machado *et al.* 2006; Amorim *et al.* 2011) and are implemented in the package. As a drawback, these alternative methods will suffer from a larger variance in estimation, particularly when the sample size is small and there is a large censoring degree. One alternative method for these scenarios was provided by Keilegom *et al.* (2011) for a progressive three-state model. The key of this methodology is the transfer of tail information from lightly censored areas to heavily ones.

The package also implements two methods that account for dependent censoring and allow for the inclusion of covariates. These two approaches are free from the Markov assumption. The functions implementing these methods use a kernel density and a bandwidth. We believe that the choice of the kernel density has relatively little impact on the estimation results. However, the use of different bandwidths might have a substantial effect on the performance of the estimators. To this end we implemented the use of the `dpik` function which is available from the R **KernSmooth** package (Wand 2014). It might be worthwhile to include other options and to investigate their impact on the estimation results.

A function called `TPmsmOut` can be used to convert an object of class '`data.frame`' with the structure of the data input as described in Section 3 to the structure of the data input used in the **p3state.msm** package. Essentially, this involves a transformation of some variables and a renaming of other variables. With this function users may connect the **TPmsm** package with the **p3state.msm** package and perform Cox-type multi-state regression.

We plan to constantly update the **TPmsm** package to improve its limitations and to cope with other estimators.

Acknowledgments

This research was financed by FEDER Funds through "Programa Operacional Factores de Competitividade – COMPETE" and by Portuguese Funds through FCT – "Fundação para a Ciência e a Tecnologia", in the form of grants PTDC/MAT/104879/2008 and PEst-OE/MAT/UI0013/2014. Thanks to the anonymous referee for comments and suggestions which have improved the presentation of the article.

References

Aalen O, Johansen S (1978). "An Empirical Transition Matrix for Non-Homogeneous Markov and Chains Based on Censored Observations." *Scandinavian Journal of Statistics*, **5**(3), 141–150. URL https://www.jstor.org/stable/4615704.

Allignol A, Latouche A (2014). "CRAN Task View: Survival Analysis." Version 2014-09-22, URL https://CRAN.R-project.org/view=Survival.

Allignol A, Schumacher M, Beyersmann J (2011). "Empirical Transition Matrix of Multi-State Models: The **etm** Package." *Journal of Statistical Software*, **38**(4), 1–15. doi:10.18637/jss.v038.i04.

Amorim AP, de Uña-Álvarez J, Meira-Machado L (2011). "Presmoothing the Transition Probabilities in the Illness-Death Model." *Statistics & Probability Letters*, **81**(7), 797–806. doi:10.1016/j.spl.2011.02.017.

Andersen PK, Borgan Ø, Gill RD, Keiding N (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Andersen PK, Keiding N (2002). "Multi-State Models for Event History Analysis." *Statistical Methods in Medical Research*, **11**(2), 91–115. doi:10.1191/0962280202SM276ra.

Andersen PK, Perme MP (2008). "Inference for Outcome Probabilities in Multistate Models." *Lifetime Data Analysis*, **14**(4), 405–431. doi:10.1007/s10985-008-9097-x.

Araújo A, Roca-Pardiñas J, Meira-Machado L (2014). ***TPmsm**: Estimation of Transition Probabilities in Multistate Models*. R package version 1.2.0, URL https://CRAN.R-project.org/package=TPmsm.

Beran R (1981). "Nonparametric Regression with Randomly Censored Survival Data." *Technical report*, University of California, Berkeley.

Borgan Ø (1998). "Aalen-Johansen Estimator." In *Encyclopedia of Biostatistics*, volume 1, pp. 5–10. John Wiley & Sons, Chichester.

Byar DP (1980). "Veterans Administration Study of Chemoprophylaxis for Recurrent Stage I Bladder Tumors: Comparisons of Placebo, Pyridoxine and Topical Thiotepa." In *Bladder Tumors and Other Topics in Urological Oncology*, volume 1 of *Ettore Majorana International Science Series*, pp. 363–370. Springer-Verlag.

Commenges D (1999). "Multi-State Models in Epidemiology." *Lifetime Data Analysis*, **5**(4), 315–327. doi:10.1023/A:1009636125294.

Cox DR, Miller HD (1965). *The Theory of Stochastic Processes*. Chapman and Hal, London.

Datta S, Satten GA (2001). "Validity of the Aalen-Johansen Estimators of Stage Occupation Probabilities and Nelson Aalen Integrated Transition Hazards for Non-Markov Models." *Statistics & Probability Letters*, **55**(4), 403–411. doi:10.1016/S0167-7152(01)00155-9.

Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, New York.

Dikta G (1998). "On Semiparametric Random Censorship Models." *Journal of Statistical Planning and Inference*, **66**(2), 253–279. doi:10.1016/S0378-3758(97)00091-8.

Ferguson N, Datta S, Brock G (2012). "**msSurv**: An R Package for Nonparametric Estimation of Multistate Models." *Journal of Statistical Software*, **50**(14), 1–24. doi:10.18637/jss.v050.i14.

Gray B (2014). "**cmprsk**: Subdistribution Analysis of Competing Risks." R package version 2.2-7, URL https://CRAN.R-project.org/package=cmprsk.

Hougaard P (1999). "Multi-State Models: A Review." *Lifetime Data Analysis*, **5**(3), 239–264. doi:10.1023/A:1009672031531.

Hougaard P (2000). *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. Springer-Verlag, New York.

Jackson CH (2011). "Multi-State Models for Panel Data: The **msm** Package for R." *Journal of Statistical Software*, **38**(8), 1–28. doi:10.18637/jss.v038.i08.

Kaplan EL, Meier P (1958). "Nonparametric Estimation From Incomplete Observations." *Journal of the American Statistical Association*, **53**(282), 457–481. doi:10.1080/01621459.1958.10501452.

Karl AT, Eubank R, Milovanovic J, Reiser M, Young D (2014). "Using RngStreams for parallel random number generation in C++ and R." *Computational Statistics*, **29**(5), 1301–1320. doi:10.1007/s00180-014-0492-3.

Keilegom IV, de Uña-Álvarez J, Meira-Machado L (2011). "Nonparametric Location-Scale Models for Successive Survival Times Under Dependent Censoring." *Journal of Statistical Planning and Inference*, **141**(3), 1118–1131. doi:10.1016/j.jspi.2010.09.010.

L'Ecuyer P (1999). "Good parameters and implementations for combined multiple recursive random number generators." *Operations Research*, **47**(1), 159–164. doi:10.1287/opre.47.1.159.

L'Ecuyer P, Simard R, Chen EJ, Kelton WD (2002). "An object-oriented random-number package with many long streams and substreams." *Operations Research*, **50**(6), 1073–1075. doi:10.1287/opre.50.6.1073.358.

Lin DY, Sun W, Ying Z (1999). "Nonparametric Estimation of the Time Distributions for Serial Events with Censored Data." *Biometrika*, **86**(1), 59–70. doi:10.1093/biomet/86.1.59.

Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C (2006). "Nonparametric Estimation of Transition Probabilities in a Non-Markov Illness-Death Model." *Lifetime Data Analysis*, **12**(3), 325–344. doi:10.1007/s10985-006-9009-x.

Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, Andersen PK (2009). "Multi-State Models for the Analysis of Time to Event Data." *Statistical Methods in Medical Research*, **18**(2), 195–222. doi:10.1177/0962280208092301.

Meira-Machado L, de Uña-Álvarez J, Somnath D (2012). "Conditional Transition Probabilities in a Non-Markov Illness-Death Model." *Discussion Papers in Statistics and Operation Research 12/05*, Universidade de Vigo. URL https://depc05.webs.uvigo.es/reports/12_05.pdf.

Meira-Machado L, Roca-Pardiñas J (2011). "**p3state.msm**: Analyzing Survival Data from an Illness-Death Model." *Journal of Statistical Software*, **38**(3), 1–18. doi:10.18637/jss.v038.i03.

Meira-Machado L, Roca-Pardiñas J, Keilegom IV, Cadarso-Suárez C (2013). "Bandwidth Selection for the Estimation of Transition Probabilities in the Location-Scale Progressive Three-State Model." *Computational Statistics*, **28**(5), 2185–2210. doi:10.1007/s00180-013-0402-0.

Moertel CG, Fleming TR, McDonald JS, *et al.* (1990). "Levamisole and Fluorouracil for Adjuvant Therapy of Resected Colon Carcinoma." *New England Journal of Medicine*, **322**(6), 352–358. doi:10.1056/NEJM199002083220602.

Moreira AC, de Uña-Álvarez J, Meira-Machado L (2013). "Presmoothing the Aalen-Johansen Estimator in the Illness-Death Model." *Electronic Journal of Statistics*, **7**, 1491–1516. doi:10.1214/13-EJS816.

Putter H (2011). "Special Issue about Competing Risks and Multi-State Models." *Journal of Statistical Software*, **38**(1), 1–4. doi:10.18637/jss.v038.i01.

Putter H, Fiocco M, Geskus R (2007). "Tutorial in Biostatistics: Competing Risks and Multi-State Models." *Statistics in Medicine*, **26**(11), 2389–2430. doi:10.1002/sim.2712.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Scheike TH, Martinussen T (2006). *Dynamic Regression Models for Survival Data*. Springer-Verlag, New York.

Scheike TH, Zhang MJ (2011). "Analyzing Competing Risk Data Using the R **timereg** Package." *Journal of Statistical Software*, **38**(2), 1–15. doi:10.18637/jss.v038.i02.

Therneau TM (2014). *A Package for Survival Analysis in S*. R package version 2.37-7, URL https://CRAN.R-project.org/package=survival.

Therneau TM, Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Wand M (2014). ***KernSmooth**: Functions for Kernel Smoothing for Wand & Jones (1995)*. R package version 2.23-13, URL https://CRAN.R-project.org/package=KernSmooth.

Affiliation:

Artur Araújo
Department of Mathematics and Applications
Centre of Mathematics
University of Minho
4810-058 Azurém, Guimarães, Portugal
E-mail: artur.stat@gmail.com

Luís Meira-Machado
Department of Mathematics and Applications
Centre of Mathematics
University of Minho
4810-058 Azurém, Guimarães, Portugal
Telephone: +351/253510400
Fax: +351/253510401
E-mail: lmachado@math.uminho.pt