Journal of Statistical Software

January 2011, Volume 38, Issue 3. *http://www.jstatsoft.org/*

p3state.msm: Analyzing Survival Data from an Illness-Death Model

Luís Meira-Machado
University of Minho

Javier Roca-Pardiñas
University of Vigo

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. In the multi-state framework, issues of interest include the study of the relationship between covariates and disease evolution, estimation of transition probabilities, and survival rates. This paper introduces **p3state.msm**, a software application for R which performs inference in an illness-death model. It describes the capabilities of the program for estimating semi-parametric regression models and for implementing nonparametric estimators for several quantities. The main feature of the package is its ability for obtaining non-Markov estimates for the transition probabilities. Moreover, the methods can also be used in progressive three-state models. In such a model, estimators for other quantities, such as the bivariate distribution function (for sequentially ordered events), are also given. The software is illustrated using data from the Stanford Heart Transplant Study.

Keywords: Kaplan-Meier estimator, Markov process, multi-state model, proportional hazards model.

1. Introduction

In many medical studies, patients may experience several events. Analysis in such studies is often performed using multi-state models (Andersen, Borgan, Gill, and Keiding 1993; Meira-Machado, Cadarso-Suárez, de Uña-Álvarez, and Andersen 2009). These models are very useful for describing event-history data, affording a better understanding of the disease process, and leading to a better knowledge of the evolution of the disease over time. Issues of interest include the estimation of transition probabilities, survival rates or assessing the effects of individual risk factors.

Although the mortality model for survival analysis can be considered the simplest multi-state model, the scope of multi-state models provides a rich framework for handling complex

Figure 1: Progressive three-state model.

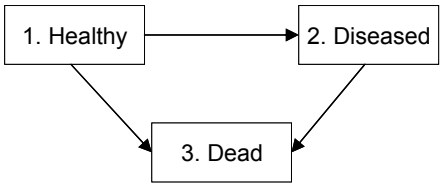

Figure 2: Illness-death model.

situations that involve more than two states and a number of possible transitions among them. The most common models in the literature include: the progressive three-state model depicted in Figure 1; and the illness-death model, also known as the disability model (Figure 2). These models can be used to study the effect of binary time-dependent covariates, such as the appearance of "recurrence" in a breast cancer study (Cadarso-Suárez, Meira-Machado, Kneib, and Gude 2010), "bleeding episodes" in patients with liver cirrhosis (Andersen, Esbjerg, and Sørensen 2000), or "transplantation" in heart diseases (Meira-Machado *et al.* 2009).

More examples of multi-state models can be found in books by Andersen *et al.* (1993) and Hougaard (2000), or in papers by Hougaard (1999) and Putter, Fiocco, and Geskus (2007).

Despite its potential, multi-state modeling is not used by practitioners as frequently as other survival analysis techniques. It is our belief that lack of knowledge of available software and non-implementation of the new methodologies in user-friendly software are probably responsible for this neglect. One important contribution to this issue was given by the R/S-PLUS **survival** package (Therneau and Lumley 2010). Thanks to this package, survival analysis is no longer limited to Kaplan-Meier curves and simple Cox models. Indeed, this package enables users to implement the methods introduced by Therneau and Grambsch (2000) for modeling multi-state survival data. In R (R Development Core Team 2010), multi-state regression can also be performed using the **msm** package by Christopher Jackson (continuous-time Markov and hidden Markov multi-state models; Jackson (2011)) and **mstate** (de Wreede, Fiocco, and Putter 2010, 2011). The **changeLOS** package Wangler, Beyersmann, and Schumacher (2006) implements the Aalen–Johansen estimator (Aalen and Johansen 1978) for general multi-state models, and the **etm** package Allignol, Schumacher, and Beyersmann (2011) has recently enabled the transition matrix to be computed, along with a covariance estimator. Meira-Machado, Cadarso-Suárez, and de Uña-Álvarez (2007) developed a software package called **tdc.msm** (available from `http://www.mct.uminho.pt/lmachado/Rlibrary`) to analyze multi-state survival data. This software may be used to fit the time-dependent Cox regression model but also several multi-state regression models in continuous time. Advantages of this software include the same data input for fitting the different models while providing the corresponding numerical and graphical outputs.

This paper describes the R-based **p3state.msm** (available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=p3state.msm`) package's capabilities for

analyzing survival data from an illness-death model. It extends existing semi-parametric regression capabilities included in many statistical software programs, such as R, S-PLUS, SAS, etc. Moreover, **p3state.msm** enables several quantities of interest to be estimated, such as transition probabilities, bivariate distribution function, etc. In addition, the current version of **p3state.msm** can also be used to draw inferences in the progressive three-state model depicted in Figure 1. This software can be used to fit the time-dependent Cox regression model (TDCM) as well as semi-parametric Cox proportional hazard regression models (Cox 1972) to all permitted transitions, by decoupling the whole process into various survival models (Cox Markov Model, CMM, and Cox semi-Markov Model, CSMM, Andersen *et al.* 2000). Numerical and graphical output for all methods can be easily obtained.

Regression (Cox-like) models can be fitted using, e.g., the **tdc.msm** software, and estimation of the Aalen-Johansen (Markov) estimator can be computed using the **etm** package. However, in the absence of the Markov property, without our software, appropriate methods for the computation of the transition probabilities would still be lacking. Moreover, in the progressive three-state model the package **p3state.msm** also provides other summary measures that greatly helps to understand the disease process.

The following section provides a brief introduction to the methodological background. Notation is introduced and nonparametric estimators for bivariate distribution function and transition probabilities are presented. Regression methods based on semi-parametric Cox regression models are also discussed. An overview of the use of **p3state.msm** is given in Section 3 and an example of its application in Section 4. A discussion is in Section 5.

2. Methodological background

Multi-state processes are characterized through transition probabilities between states h and j, which are expressed for $s \leq t$ as

$$p_{hj}(s,t) = p\left(X(t) = j | X(s) = h, H_{s^-}\right)$$

where $H_{s^-}(\sigma - algebra)$ denotes the history of the process, which is generated and consists of the observation of the process over the interval $[0, s)$; or through transition intensities, which are expressed as

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

representing the instantaneous hazard of progression to state j conditionally on occupying state h, and which are assumed to exist.

A number of possible models for the transition rates have been studied. These include: time-homogeneity; the Markov assumption; and the semi-Markov assumption.

This section will focus on the estimation of transition probabilities for the illness-death model (Meira-Machado, de Uña-Álvarez, and Cadarso-Suárez 2006). These estimators also apply to the case of the progressive three-state model, since this model can be viewed as a particular case of the illness-death model where no transitions are observed on disease-free mortality transition ($1 \to 3$). For the progressive three-state model, estimators are also derived for the bivariate distribution function of the pair of gap times and for the distribution of the second gap time (de Uña-Álvarez and Meira-Machado 2008). The idea behind the proposed estimators is using the Kaplan-Meier estimator pertaining to the distribution of

the total time to weight the data. The estimators for the transition probabilities can be regarded as an alternative to Aalen-Johansen estimators since they do not rely on the Markov assumption.

2.1. The illness-death model

Consider the illness-death model depicted in Figure 2. Let $\{X(t), t \geq 0, X(0) = 1\}$ denote the underlying stochastic process, where $X(t)$ denotes the state being occupied at time t, for which all individuals are in state 1 at time zero. 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. The survival time of the process is given by $T = I(T_{12} \leq T_{13})(T_{12} + T_{23}) + I(T_{12} > T_{13})(T_{13})$.

This random vector may be subjected to a random right-censoring variable, denoted as C and assumed to be independent of (T_{12}, T_{13}, T_{23}) . Owing to censoring, only the following are observed: sojourn time in state 1, $U = \min(T_{12}, T_{13}, C)$; sojourn time in state 2, $V = \min(T_{23}, C - T_{12})$; observed total time $Y = U + \delta V = \min(T, C)$ $(\delta = I\,(T_{12} \leq \min(T_{13}, C)))$; and indicator statuses $\Delta_1 = I\,(\min(T_{12}, T_{13}) \leq C)$ and $\Delta_2 = I(T \leq C)$.

Traditionally, the transition probabilities are estimated via the non-parametric model (Aalen-Johansen estimator, Aalen and Johansen 1978). The performance of Aalen-Johansen estimator of stage occupancy probabilities was investigated by Datta and Satten (2001) when the process is not Markovian. These authors concluded that the Aalen-Johansen method provides consistent estimators in this case. However, no similar result is available when the target is the transition (rather than the occupancy) probability. Recently, Meira-Machado *et al.* (2006) verified that, in non-Markov situations, the use of these estimators for empirical estimation of the transition probabilities, $p_{hj}(s, t)$, may be inappropriate. Within the scope of the illness-death model, these authors propose alternative estimators for the transition probabilities, which do not rely on the Markov assumption. The quantities are determined by the joint distribution of (T_{12}, T_{13}, T_{23}). Specifically, knowledge of the distribution H of $\min(T_{12}, T_{13})$ will suffice for recovery of $p_{11}(s, t)$ while expectations of type $S(\phi) = E[\phi(U, Y)]$ arise when handling $p_{12}(s, t)$ and $p_{22}(s, t)$. The estimators are expressed as

$$\hat{p}_{11}(s,t) = \frac{1 - \hat{H}(t)}{1 - \hat{H}(s)} \tag{1}$$

$$\hat{p}_{12}(s,t) = \frac{\sum_{i=1}^{n} W_i \phi_{s,t}\left(U_{[i]}, Y_{(i)}\right)}{1 - \hat{H}(s)} \tag{2}$$

$$\hat{p}_{22}(s,t) = \frac{\sum_{i=1}^{n} W_i \tilde{\phi}_{s,t}\left(U_{[i]}, Y_{(i)}\right)}{\sum_{i=1}^{n} W_i \tilde{\phi}_{s,s}\left(U_{[i]}, Y_{(i)}\right)} \tag{3}$$

where W_i are the Kaplan-Meier weights attached to $Y_{(i)}$, $\hat{H}$ is the Kaplan-Meier estimator based on the pairs (U_i, Δ_{1i}) , and $\phi_{s,t}(u, v) = I(s < u \leq t, v > t)$ and $\tilde{\phi}_{s,t}(u, v) = I(u \leq s, v > t)$. In these expressions, $Y_{(1)} \leq \ldots \leq Y_{(n)}$ denotes the ordered sample of the Y_i's and $U_{[i]}$ for the pair attached (concomitant) to the $Y_{(i)}$ value.

Note that for the illness-death model, the transition probabilities to be estimated reduce to $p_{11}(s, t)$, $p_{12}(s, t)$, and $p_{22}(s, t)$, since $p_{13}(s, t) = 1 - p_{11}(s, t) - p_{12}(s, t)$ and $p_{23}(s, t) = 1 - p_{22}(s, t)$.

2.2. The progressive three-state model

Consider the progressive three-state model depicted in Figure 1. The stochastic behavior of the process is represented by a random vector (T_{12}, T_{23}), deemed to be a pair of gap times of successive events which may be subjected to random right-censoring. Let C be the right-censoring variable, assumed to be independent of (T_{12}, T_{23}), and let $T = T_{12} + T_{23}$ be the survival time of the process and $Y = \min(T, C)$ the observed total time.

The methods shown above (for the illness-death model, 1–3) apply to the progressive three-state model. Estimation of the marginal distribution for the second gap time, $F_2(y) = P(T_{23} \leq y)$, and estimation of the bivariate distribution function, $F_{12}(x, y) = P(T_{12} \leq x, T_{23} \leq y)$, constitute two additional topics of interest within the scope of the model depicted in Figure 1. Indeed, as T_{23} and $C_2 = (C - T_{12}) I(T_{12} \leq C)$ will in general be interdependent, estimation of the marginal distribution for the second gap time is not a simple issue. The same applies to the bivariate distribution function. A simple estimator was proposed by de Uña-Álvarez and Meira-Machado (2008), with the Kaplan-Meier estimator pertaining to the distribution of the total time being used to weight the data. The estimators are expressed as

$$\hat{F}_{12}(x, y) = \sum_{i=1}^{n} W_i I(U_i \leq x, V_i \leq y) \tag{4}$$

where W_i is the Kaplan-Meier weight attached to Y_i when estimating the marginal distribution of Y from (Y_i, Δ_{2i})s, $U_i = \min(T_{12i}, C_i)$ and $V_i = \min(T_{23i}, C_{2i})$. From (4) one can obtain an estimator for the marginal distribution of the second gap time, namely, $\hat{F}_2(y) = \hat{F}_{12}(\infty, y) = \sum_{i=1}^{n} W_i I(V_i \leq y)$.

2.3. Regression models

One important goal in multi-state modeling is to study the relationships between the different predictors and the outcome. To relate the individual characteristics to the intensity rates through a possibly time-dependent covariate vector, Z, several models have been used in literature. A common simplifying strategy is to decouple the whole process into various survival models, by fitting separate intensities to all permitted transitions using semi-parametric Cox proportional hazard regression models, while making appropriate adjustments to the risk set. For the illness-death model of Figure 2, the transition intensities, $\alpha_{hj}(t; Z)$, $1 \leq h < j \leq 3$, may be modeled using Cox-like models of the form $\alpha_{hj}(t; Z) = \alpha_{hj0}(t) \exp\left(\beta_{hj}^T Z\right)$ assuming the process to be Markovian. These models are known as Cox Markov models (CMM). The Markov assumption states that the future depends on the individual's past solely by means of his current state. However, by ignoring disease history behavior, Markov models may have severe limitations, thus rendering them inappropriate. One alternative approach is to use a Cox semi-Markov model (CSMM) in which the future of the process does not depend on the current time but rather on the duration in the current state. These models are also called "clock reset" models, because each time the patient enters a new state time is reset to 0. Assuming an illness-death model, the only difference between CMM and CSMM (Andersen *et al.* 2000) resides in transition $2 \to 3$, in which intensity α_{23} is modeled in a different way. Specifically, the corresponding intensity α_{23} in the CSMM is given by $\alpha_{23}(t - T_{12}; Z) = \alpha_{230}(t - T_{12}) \exp\left(\beta_{23}^T Z\right)$ where T_{12} is the entry time into state 2. These

Cox-like models (Markovian and semi-Markovian) can be fitted by means of most of the statistical packages (R, S-PLUS, SAS, etc.), provided that a counting process notation is used, with each patient being represented by several observations (Meira-Machado *et al.* 2009).

Though we assume different baseline hazards for the transition intensities, we note that, in some cases, one may impose some restriction on the baseline hazards. For example, for the illness-death model, one approach that is often considered is to assume the baseline hazards for transition $1 \to 3$ and for the $2 \to 3$ transition to be proportional. In such cases, the model for these transitions is given by $\alpha_{13}(t; Z) = \alpha_{130}(t) \exp\left(\beta_{13}^T Z\right)$ and $\alpha_{23}(t; Z) = \alpha_{130}(t) \exp\left(\beta_{23}^T Z + \delta\right)$.

3. Package description

The **p3state.msm** software contains nonparametric statistical methods for estimating quantities of interest such as transition probabilities, the bivariate distribution function for censored gap times, etc. This software is intended to be used with the R statistical program (R Development Core Team 2010). In the R language, programming is based on objects, and computations are basically specialized functions designed to perform specific calculations. Our package is composed of 6 functions that enable users to fit the proposed models and methods. Table 1 provides a summary of the objects in this package.

Users can fit the proposed models and methods discussed in the previous section by means of the three functions, namely, `p3state`, `summary` and `plot`. Table 2 provides a summary of the arguments in the three functions.

It should be noted that only `data` is a required argument. Records in the data file must contain the following variables: `times1`, `delta`, `times2`, `time`, `status`, `covariate1`, `covariate2`, and so on. The structure of the data input is as follows: each individual is represented by one line

Function	Description
`p3state`	Main function for fitting regression models and obtaining multi-state estimates (transition probabilities, bivariate distribution function, etc.).
`plot`	A function that provides the plots for transition probabilities, bivariate distribution function, and marginal distribution of the second time (the last two available solely for the progressive three-state model).
`summary`	Summary method for objects of class `p3state`.
`data.creation.reg`	Provides the adequate dataset for implementing regression models (TDCM, CMM, and CSMM). Same input data as for `p3state`.
`pLIDA`	Provides estimates for the transition probabilities using the methods in paper by Meira-Machado *et al.* (2006).
`Biv`	Provides estimates for the bivariate distribution function, using the paper by de Uña-Álvarez and Meira-Machado (2008). Available solely for the progressive three-state model.

Table 1: Summary of functions in the package.

p3state arguments	Description
data	Data input as described below.
coxdata	Data set in a counting process data-structure. This data set can be obtained by using data.creation.reg. If NULL, the main function p3state function will automatically create this dataset every time it is called.
formula	A formula giving the vector of covariates, e.g., formula = ~ age + sex.
regression	A logical variable indicating whether you want the regression model (from the Cox regression model and Cox-type multi-state models).
summary arguments	**Description**
object	Object of class p3state.
model	A character string specifying which model(s) to fit. Possible values are TDCM, CMM and CSMM. If NULL, none of the regression models will be implemented.
covmat	If TRUE, provides the variance-covariance matrices when implementing regression models. By default, covmat = FALSE.
estimate	A logical variable indicating whether you want the nonparametric estimates from the multi-state model. These include transition probabilities, bivariate distribution function, and marginal distribution of the second time (the last two available solely for the progressive three-state model).
time1	The first time for computing estimates for the transition probabilities and bivariate distribution function. NULL is equivalent to 0.
time2	The second time for computing estimates for the transition probabilities and bivariate distribution function.
plot arguments	**Description**
x	Object of class p3state.
plot.trans	A character string specifying which plot(s) are to be given for the transition probabilities. Possible values are all, P11, P12, P22 and P23.
plot.marginal	If TRUE, plots the marginal distribution of the second gap time.
plot.bivariate	If TRUE, plots the bivariate distribution function.
time1	Starting value for computing the transition probabilities. NULL is equivalent to 0.
col.biv	A logical variable indicating whether you want color to be used in the filled.contour plot. By default col.biv = FALSE.

Table 2: Summary of arguments of the p3state, summary, and plot functions.

of data. The variable times1 represents the observed time in state 1, and delta the indicator of transition to state 2 (taking a value of 1 if a transition to state 2 is observed, and a value of 0 otherwise). The variable times2 represents the observed time in state 2. If no transition into state 2 (delta = 0) is observed then times2 = 0. The variable time is just the observed total time (times1 + times2) whereas status is the final status of the individual (1 if the

event of interest, representing state 3, is observed and 0 otherwise). The following variables are the covariates to be studied in the regression models. Note that possible courses for the individual include: $1 \to 1$ (the individual remains in state 1 until the end of the study; if `delta = 0` and `status = 0`); $1 \to 3$ (a direct transition from state 1 into state 3 is observed; if `delta = 0` and `status = 1`); $1 \to 2 \to 2$ (if `delta = 1` and `status = 0`); and $1 \to 2 \to 3$ (if `delta = 1` and `status = 1`).

Although only `data` is a required argument in `p3state`, note that for implementing regression models (alone) the argument `formula` is also necessary. The `p3state` function returns an object of class `p3state` with the following components:

- `descriptives`: vector with transition between states.
- `datafr`: `data.frame` to be used for obtaining the nonparametric estimates and plotting purposes.
- `tdcm`: a `coxph` object with the fit of the Cox regression model with time-dependent covariates.
- `msm12`: a `coxph` object with the fit of the Cox model for transition from state 1 to state 2.
- `msm13`: a `coxph` object with the fit of the Cox model for transition from state 1 to state 3 (solely for the illness-death model).
- `cmm23`: a `coxph` object with the fit of the Cox Markov model for transition from state 2 to state 3.
- `csmm23`: a `coxph` object with the fit of the Cox semi-Markov model for transition from state 2 to state 3.
- `tma`: a `coxph` object with the fit of a Cox model for testing the Markov assumption.
- `tma2`: the same as `tma` but with all the covariates in the model.

The object obtained when using the `p3state` function is the only argument required for `summary`. However, the arguments, `regression`, `time1` and `time2` are also required if the results from the regression model and the estimates for the other nonparametric methods are sought. This function prints several numerical results on the screen, i.e., parameter estimates with standard errors for the covariates for TDCM, CMM and CSMM models, transition probabilities estimates, and estimates for the bivariate distribution function and the marginal distribution of the second time (only in the case of the progressive three-state model).

The `plot` function provides the following graphical output: transition probabilities estimates; bivariate distribution function; and marginal distribution of the second time.

The dataset included in **p3state.msm** package is the well-known Stanford Heart Transplant data in a different format. Details about this dataset are given below.

4. Example of application: Stanford Heart Transplant data

An example of application is provided using the Stanford Heart Transplant data. A copy of the data may be obtained from statlib or in the R **survival** package. This data set is

also available in the book by Kalbfleisch and Prentice (1980, Appendix I, pp. 230–232) or in the paper by Crowley and Hu (1977). The data set covers the period until 1974-04-01. In this period some patients died before an appropriate heart could be found. Of the 103 patients, 69 received heart transplants; the number of deaths was 75; the remaining 28 patients contributed with censored survival times. For each individual, an indicator of final vital status (censored or not), survival times (time since acceptance into the transplantation program to transplant and to death) from patient entry into the study (in days), and a vector of covariates including age at acceptance (age), year of acceptance (year), previous surgery (surgery: coded as 1 = yes; 0 = no), and transplant (coded as 1 = yes; 0 = no) were recorded. The covariate "transplant" is the only time-dependent covariate, while the other covariates included are fixed. These time-dependent covariates can be re-expressed as a multi-state model, with states based on the values of the covariate. If all subjects observe the intermediate event, then the time-dependent covariate renders it possible for the progressive three-state model to be used (Figure 1); otherwise, it is feasible for the illness-death model, depicted in Figure 2, to be used. For the transplant heart data, the time-dependent covariate can be expressed as an intermediate event that can be modeled using an illness-death model with states, "alive without transplant", "alive with transplant", and "dead". This relationship will be used below to compare the Cox model with time-dependent covariates against common multi-state regression approaches (CMM and CSMM). Other targets include the estimation of transition probabilities. This will be done using **p3state.msm**.

In the following, we will demonstrate the package capabilities using data from the Stanford Heart Transplant Study. Bellow is an excerpt of the data.frame with one row per individual

```
R> library("p3state.msm")
R> data("heart2")
R> head(heart2)
```

```
  times1 delta times2 time status        age      year surgery
1     50     0      0   50      1 -17.155373 0.1232033       0
2      6     0      0    6      1   3.835729 0.2546201       0
3      1     1     15   16      1   6.297057 0.2655715       0
4     36     1      3   39      1  -7.737166 0.4900753       0
5     18     0      0   18      1 -27.214237 0.6078029       0
6      3     0      0    3      1   6.595483 0.7008898       0
```

Individuals represented in lines 1, 2, 5 and 6 experienced a direct transition from state 1 to state 3 ($1 \to 3$); individuals represented in lines 3 and 4 had a heart transplant at 1 and 36 days, respectively, after enrolment, and died at 16 and 39 days ($1 \to 2 \to 3$), respectively. We note that `delta = 1` and `status = 0` corresponds to individuals with a transition from state 1 to state 2 and, afterwards, he/she exhibits a censored sojourn time in state 2 ($1 \to 2 \to 2$; individuals that receive a new heart and remain alive until de end of study); finally, `delta = 0` and `status = 0` corresponds to a censored sojourn time in state 1 ($1 \to 1$; remained alive without a heart transplant).

Two central questions that arise in multi-state survival data are: what is the relationship between the different covariates and disease evolution; what is the rate (hazard) at which

persons in state h move to state j. Both questions can be answered using our software. This will be shown below.

The **p3state.msm** software enables several semi-parametric Cox models to be fitted. The time-dependent Cox model or multi-state Cox-like models (CMM and CSMM) can be constructed with the following input command:

```
R> obj1.p3state <- p3state(heart2, formula = ~ age + year + surgery)
```

Results are printed on the screen as follows: For the TDCM (where `treat` denotes the time-dependent covariate), by using

```
R> summary(obj1.p3state, model = "TDCM")

  ***** TIME-DEPENDENT COX REGRESSION MODEL *****
n=  172
               coef exp(coef)   se(coef)           z   Pr(>|z|)
age      0.02716664 1.0275390 0.01371412  1.98092553 0.04759963
year    -0.14634635 0.8638585 0.07046798 -2.07677794 0.03782206
surgery -0.63720989 0.5287657 0.36722600 -1.73519821 0.08270570
treat   -0.01025077 0.9898016 0.31375480 -0.03267128 0.97393672

        exp(coef) exp(-coef) lower .95 upper .95
age     1.0275390   0.973199 1.0002875 1.0555330
year    0.8638585   1.157597 0.7524197 0.9918021
surgery 0.5287657   1.891197 0.2574423 1.0860419
treat   0.9898016   1.010303 0.5351550 1.8306980

Likelihood ratio test=  15.11148 on  4  df, p= 0.0044755

-2*Log-likelihood= 581.1312
```

here `treat` denotes the time-dependent covariate associated with the occurrence of the intermediate state (in our application, transplantation, which is a binary time-dependent covariate); and in which the likelihood ratio test is of the model with all covariates versus a model with intercept only.

Multi-state Cox-like models (CMM and CSMM) can be obtained by simply changing the `model` argument to "CMM" or "CSMM", e.g.,

```
R> summary(obj1.p3state, model = "CMM")

********************** COX MARKOV MODEL **********************
    *************** FROM STATE 1 TO STATE 3 ****************
n=  103
               coef exp(coef)   se(coef)          z   Pr(>|z|)
age      0.01978539 1.0199824 0.01807908  1.0943806 0.27378810
year    -0.28331015 0.7532861 0.11096315 -2.5531913 0.01067409
surgery -0.22875449 0.7955238 0.63608541 -0.3596286 0.71912491
```

```
        exp(coef) exp(-coef) lower .95 upper .95
age     1.0199824   0.980409 0.9844729 1.0567728
year    0.7532861   1.327517 0.6060493 0.9362934
surgery 0.7955238   1.257033 0.2286737 2.7675156

Likelihood ratio test=  8.623363 on  3  df, p= 0.03474115
-2*Log-likelihood= 214.9848

    *************** FROM STATE 1 TO STATE 2 ***************
n=  103
                coef exp(coef)   se(coef)          z   Pr(>|z|)
age     0.0311147186  1.031604 0.01398119 2.22546929 0.02604975
year    0.0007505999  1.000751 0.06948591 0.01080219 0.99138127
surgery 0.0473360792  1.048474 0.31524102 0.15015838 0.88063966

        exp(coef) exp(-coef) lower .95 upper .95
age      1.031604  0.9693644 1.0037190  1.060263
year     1.000751  0.9992497 0.8733322  1.146760
surgery  1.048474  0.9537668 0.5652286  1.944874

Likelihood ratio test=  5.768582 on  3  df, p= 0.1234284
-2*Log-likelihood= 509.5638

    *************** FROM STATE 2 TO STATE 3 ***************
n=  69
               coef exp(coef)   se(coef)          z   Pr(>|z|)
age      0.04956295 1.0508117 0.02137741  2.3184737 0.02042359
year    -0.02303487 0.9772284 0.09693819 -0.2376243 0.81217248
surgery -0.81647952 0.4419849 0.45491690 -1.7947883 0.07268744

        exp(coef) exp(-coef) lower .95 upper .95
age     1.0508117  0.9516452 1.0076935  1.095775
year    0.9772284  1.0233022 0.8081317  1.181708
surgery 0.4419849  2.2625206 0.1812097  1.078036

Likelihood ratio test=  11.30435 on  3  df, p= 0.01018901
-2*Log-likelihood= 290.1922

Checking the Markov assumption:
Testing if the time spent in state 1 (start) is important on transition from
state 2 to state 3

              coef exp(coef)    se(coef)         z   Pr(>|z|)
start -0.009392569 0.9906514 0.005340591 -1.758713 0.07862619

The p-value is  0.07862619
```

Note that differences between the CMM and CSMM are only present in the transition from state 2. The results reported above show that Markov's assumption is satisfactory for the Stanford Heart data. This assumption is tested by including covariates depending on the history, "time spent in state 1" (Kay 1986). Note that rather than "time spent in state 1" one could use time-dependent indicator covariates (Andersen *et al.* 2000); or test if "time since transplant" (Hougaard 1999) is important in mortality transition after transplantation. We note that this assumption can also be performed considering the model with all covariates. The output for such model can be obtained using the command `obj1.p3state$tma2` (results not shown).

The patients course over time may also be studied through transition probabilities. To obtain these estimates (for a model with no covariates), the following input command must be typed:

```
R> summary(obj1.p3state, time1 = 20, time2 = 200)

Illness-death model

The estimate of the transition probability P11( 20 , 200 ) is  0.1040599
The estimate of the transition probability P12( 20 , 200 ) is  0.2821365
The estimate of the transition probability P13( 20 , 200 ) is  0.6138035
The estimate of the transition probability P22( 20 , 200 ) is  0.3540728
The estimate of the transition probability P23( 20 , 200 ) is  0.6459272
```

The results obtained with the last two input commands can be obtained with a single input command (results not shown), namely:

```
R> summary(obj1.p3state, model = "CMM", time1 = 20, time2 = 200)
```

The package also provides plots for several functions. Estimates of the the transition probabilities (for a model with no covariates) are displayed in Figure 3. These plots can be obtained with:

```
R> plot(obj1.p3state, plot.trans = "all", time1 = 20)
```

Just for the purposes of illustration, we created a new data set in which the individuals who experienced a direct transition from state 1 to state 3 ($1 \rightarrow 3$) are taken as censored on death time. The aim of this subset is to illustrate the program in a progressive three-state model. This can be done with the following three lines of command:

```
R> p <- which(heart2$delta == 0 & heart2$status == 1)
R> inputdata <- heart2
R> inputdata[p,5] <- 0
```

Estimates for the transition probabilities and estimates of regression effects can be obtained in the same way as for the illness-death model.

The outputs for the bivariate distribution function and for the marginal distribution of the second gap time (time since transplantation) are useful displays that greatly helps to understand the patients' course over time. Estimates and plots for these quantities can easily be obtained. The following three input commands provide the corresponding numerical and graphical output (Figures 4, 5 and 6):

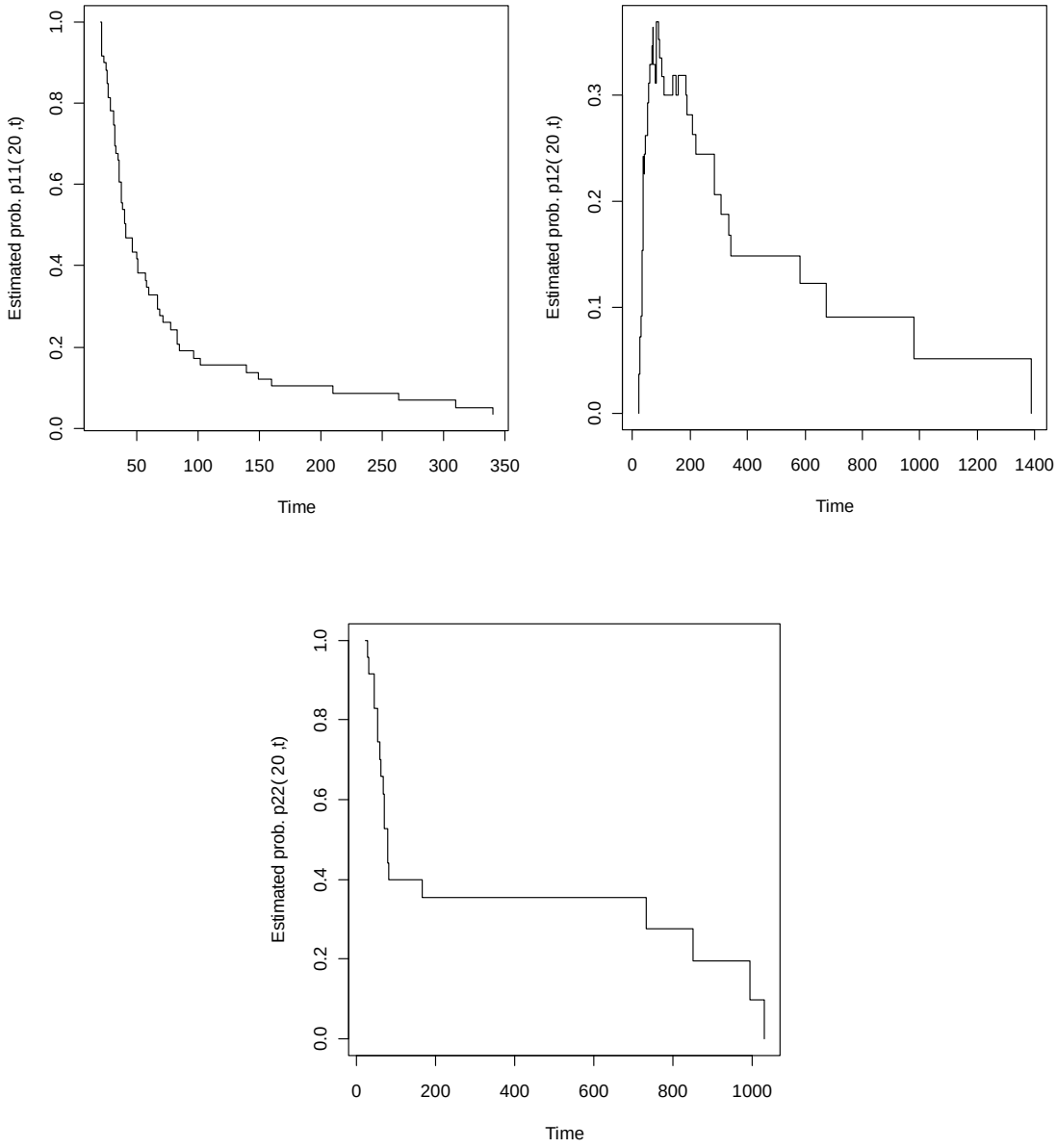

Figure 3: Transition probability estimates with first time equal to 20 days.

```
R> obj2.p3state <- p3state(inputdata)
R> summary(obj2.p3state, time1 = 50, time2 = 300)

Progressive three-state model

Number of individuals experiencing the intermediate event:  69
Number of events for the direct transition from state 1 to state 3:  0
Number of individuals remaining in state 1:  34
Number of events on transition leaving state 2:  45
Number of censored observations on transition leaving state 2:  24
```

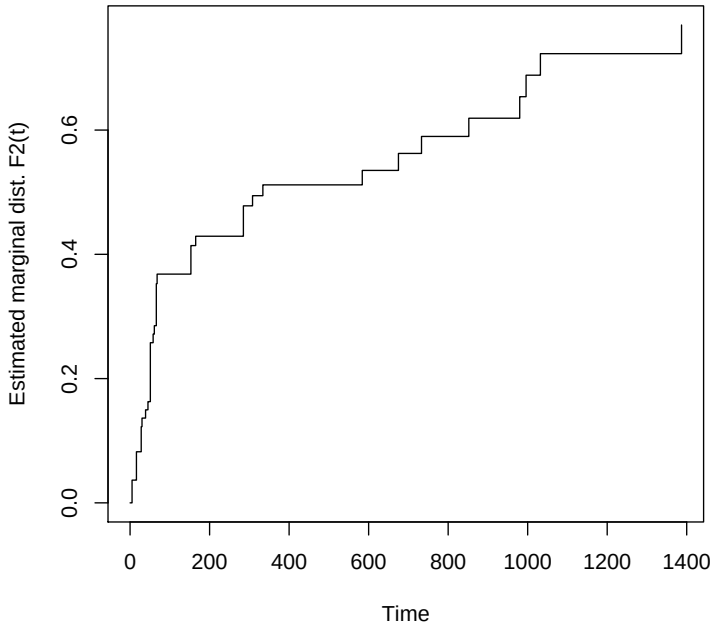

Figure 4: Marginal distribution of the second time.

```
The estimate of the transition probability P11( 50 , 300 ) is  0.3066378
The estimate of the transition probability P12( 50 , 300 ) is  0.1654678
The estimate of the transition probability P13( 50 , 300 ) is  0.5278944
The estimate of the transition probability P22( 50 , 300 ) is  0.4947382
The estimate of the transition probability P23( 50 , 300 ) is  0.5052618
The estimate of the bivariate distribution function
F12( 50 , 300 ) is  0.3899876
The estimate of the marginal distribution function of the second gap time,
F2( 300 ) is 0.4943689
```

```
R> plot(obj2.p3state, time1 = 50, plot.marginal = TRUE,
+    plot.bivariate = TRUE)
```

5. Conclusion

This paper discusses implementation in R of some newly developed methods in multi-state models. The **p3state.msm** package uses methods proposed by Meira-Machado *et al.* (2006) (transition probabilities) and de Uña-Álvarez and Meira-Machado (2008) (bivariate distribution function for the censored gap times in the progressive three-state model). The main novelty of these estimators is that they do not rely on the Markov assumption, typically assumed to hold in a multi-state model. The software also enables the user to easily ob-

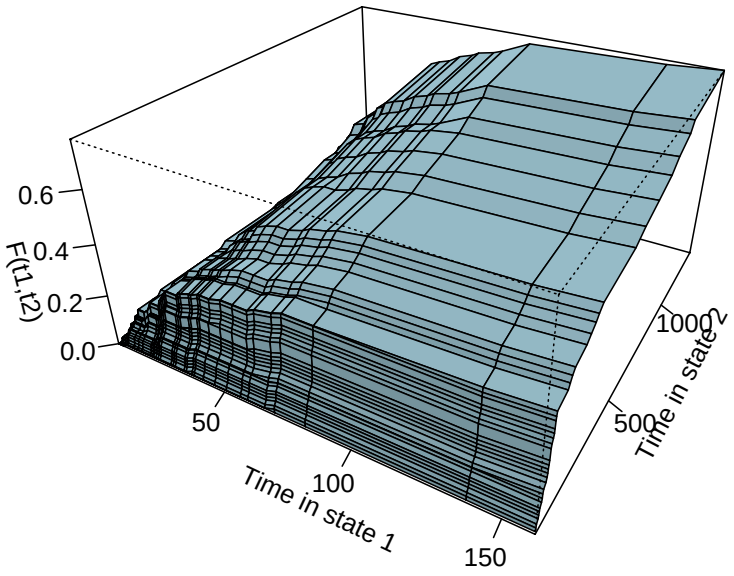

Figure 5: Bivariate distribution.

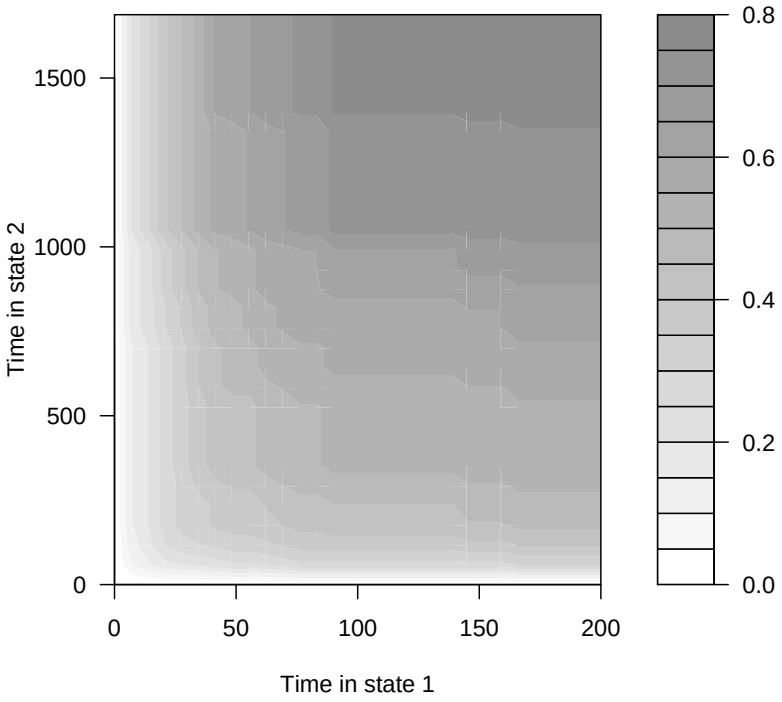

Figure 6: Contour plot for the bivariate distribution.

tain estimates of regression parameters, assuming that each transition may be specified by a Cox-type model. Numerical results as well as graphics are easily obtained.

As in other Cox analyses some care is necessary when performing regression modeling (assumptions, baseline hazards, etc.). We note that the software returns an object of class `p3state` with components that allow users to perform a detailed analysis of these models.

We mention three important topics that we shall consider in future versions of the package. First, covariates have not been included in our nonparametric methods, e.g. transition probabilities. Another topic of much practical interest is that of providing pointwise confidence bands for the transition probabilities. To this end, we note that the $(1-\alpha)100\%$ limits for the confidence interval of $p_{hj}(s,t)$ can be obtained using pointwise confidence bands based on the bootstrap. Though this can be achieved using the current version of our paper it will be quite demanding. Finally, it is our belief that it may be valuable to include an option where the baseline intensities (for Cox-like regression models) are estimated using methods described in paper by Meira-Machado *et al.* (2006).

We plan to constantly update **p3state.msm** to cope with other multi-state models such as the progressive k-state model and the bivariate model (for bivariate failure times).

Acknowledgments

The authors acknowledge receiving financial support from the Spanish Ministry of Education & Science in the form of grant MTM2008-03129 and PGIDIT07PXIB300191PR of the Xunta de Galicia. Luis F. Meira-Machado also acknowledges financial support by Grant PTDC/MAT/104879/2008 (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education, and by CMAT and FCT under the POCI 2010 program. Thanks to two anonymous referees for comments and suggestions which have improved the presentation of the article.

References

Aalen OO, Johansen S (1978). "An Empirical Transition Matrix for Nonhomogeneous Markov Chains Based on Censored Observations." *Scandinavian Journal of Statistics*, **5**, 141–150.

Allignol A, Schumacher M, Beyersmann J (2011). "Empirical Transition Matrix of Multistate Models: The **etm** Package." *Journal of Statistical Software*, **38**(4), 1–15. URL http://www.jstatsoft.org/v38/i04/.

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

Andersen PK, Esbjerg S, Sørensen TI (2000). "Multi-State Models for Bleeding Episodes and Mortality in Liver Cirrhosis." *Statistics in Medicine*, **19**, 587–599.

Cadarso-Suárez C, Meira-Machado L, Kneib T, Gude F (2010). "Flexible Hazard Ratio Curves for Continuous Predictors in Multi-State Models: A P-Spline Approach." *Statistical Modelling*, **10**(3), 291–314.

Cox DR (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society B*, **34**, 187–220.

Crowley J, Hu M (1977). "Covariance Analysis of Heart Transplant Survival Data." *Journal of the American Statistical Association*, **72**, 27–36.

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 and Probability Letters*, **55**, 403–411.

de Uña-Álvarez J, Meira-Machado LF (2008). "A Simple Estimator of the Bivariate Distribution Function for Censored Gap Times." *Statistics and Probability Letters*, **78**, 2440–2445.

de Wreede LC, Fiocco M, Putter H (2010). "The **mstate** Package for Estimation and Prediction in Non- and Semi-Parametric Multi-State and Competing Risks Models." *Computer Methods and Programs in Biomedicine*, **99**, 261–274.

de Wreede LC, Fiocco M, Putter H (2011). "**mstate**: an R Package for the Analysis of Competing Risks and Multi-State Models." *Journal of Statistical Software*, **38**(7), 1–30. URL http://www.jstatsoft.org/v38/i07/.

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

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

Jackson C (2011). "Multi-State Models for Panel Data: The **msm** Package for R." *Journal of Statistical Software*, **38**(8), 1–29. URL http://www.jstatsoft.org/v38/i08/.

Kalbfleisch JD, Prentice RL (1980). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Kay R (1986). "A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies." *Biometrics*, **42**, 855–865.

Meira-Machado L, Cadarso-Suárez C, de Uña-Álvarez J (2007). "**tdc.msm**: An R Library for the Analysis of Multi-State Survival Data." *Computer Methods and Programs in Biomedicine*, **86**, 131–140.

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

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**, 325–344.

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

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Therneau T, Lumley T (2010). ***survival***: *Survival Analysis Including Penalised Likelihood*. R package version 2.36-1, URL http://CRAN.R-project.org/package=survival.

Therneau TM, Grambsch PM (2000). *Modeling Survival Data. Extending the Cox Model*. Statistics for Biology and Health. Springer-Verlag, New York.

Wangler M, Beyersmann J, Schumacher M (2006). "**changeLOS**: An R-Package for Change in Length of Hospital Stay Based on the Aalen-Johansen Estimator." *R News*, **6**(2), 31–35. URL http://CRAN.R-project.org/doc/Rnews/.

Affiliation:

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

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 38, Issue 3 *Submitted:* 2009-06-30
January 2011 *Accepted:* 2010-03-02