Journal of Statistical Software

August 2015, Volume 66, Issue 6. *http://www.jstatsoft.org/*

SemiMarkov: An R Package for Parametric Estimation in Multi-State Semi-Markov Models

Agnieszka Król
Université de Bordeaux

Philippe Saint-Pierre
Université Pierre et Marie Curie Paris

Abstract

Multi-state models provide a relevant tool for studying the observations of a continuous-time process at arbitrary times. Markov models are often considered even if semi-Markov are better adapted in various situations. Such models are still not frequently applied mainly due to lack of available software. We have developed the R package **SemiMarkov** to fit homogeneous semi-Markov models to longitudinal data. The package performs maximum likelihood estimation in a parametric framework where the distributions of the sojourn times can be chosen between exponential, Weibull or exponentiated Weibull. The package computes and displays the hazard rates of sojourn times and the hazard rates of the semi-Markov process. The effects of covariates can be studied with a Cox proportional hazards model for the sojourn times distributions. The number of covariates and the distribution of sojourn times can be specified for each possible transition providing a great flexibility in a model's definition. This article presents parametric semi-Markov models and gives a detailed description of the package together with an application to asthma control.

Keywords: multi-state semi-Markov models, parametric estimation, exponentiated Weibull distribution, asthma, R.

1. Introduction

In multi-state models of longitudinal data usually a process is assumed to be Markovian that is that the conditional probability distribution of future states depends only on the present state, not on the whole sequence of past events. In a discrete-time framework, one can study models based on Markov chains. For instance, the R (R Core Team 2015) package **VLMC** (Maechler 2015) can be used to fit a variable length Markov chain to a discrete time series. In a continuous-time framework, multi-state models based on Markov processes are often considered. In various applications, the intensities between states are supposed to be constant

in time (homogeneity assumption) or piecewise constant (Huszti, Abrahamowicz, Alioum, Binquet, and Quantin 2012; Saint-Pierre, Combescure, Daurès, and Godard 2003; Aguirre-Hernàndez and Farewell 2002). A few R packages have been developed to simplify the usage of multi-state Markov models. The **msm** package (Jackson 2011) allows to fit homogeneous Markov or hidden Markov models in continuous-time and discrete-time. Non and semi parametric estimation of non homogeneous Markov models or competing risks models are possible using the **mstate** package (de Wreede, Fiocco, and Putter 2011). The **etm** package (Allignol, Schumacher, and Beyersmann 2011) computes the Aalen-Johansen empirical transition matrix whereas **p3state** (Meira-Machado and Pardiñas 2011) focuses on the illness-death model.

A non homogeneous Markov model is well adapted when the process evolution depends on calendar time, age or time since the beginning of the study. However, the memoryless property implies that the waiting times distributions in a Markov model is exponential. In cases when this assumption is too restrictive, semi-Markov models can be considered since they involve distributions of sojourn times as parameters. From a theoretical point of view, several results are given in Limnios and Oprisan (2001). Non homogeneous semi-Markov models are very complex and are rarely used in practical situations (Monteiro, Smirnov, and Lucas 2006). In the homogeneous case, a non parametric estimation of the semi-Markov process hazard rate can be found in Gill (1980) or in Ouhbi and Limnios (1999). The parametric maximum likelihood estimation is based on a parametric definition of the sojourn times distributions (Pérez-Ocón, Ruiz-Castro, and Gàmiz-Pérez 1999). Indeed, the Weibull or the exponentiated (generalized) Weibull distributions are efficient and flexible to fit the $\cap$ or $\cup$ shape (of the hazard rates) common in biology (Foucher, Mathieu, Saint-Pierre, Durand, and Daurès 2005), the life sciences and reliability. Moreover, the parametric model allows to incorporate covariates in the distribution of sojourn times using a proportional-hazards regression model (Cox 1972).

Few R packages have been developed to handle semi-Markov or hidden semi-Markov models. The **mhsmm** package (O'Connell and Højsgaard 2011) performs estimation and prediction for multiple observation sequences in hidden semi-Markov models. The **msSurv** package (Ferguson, Datta, and Brock 2012) provides non parametric estimation in semi-Markov models but covariates are not considered. However, it seems that the parametric approach is not yet implemented in statistical software. We have developed an R package named **SemiMarkov** (Listwon-Krol and Saint-Pierre 2015) which performs parametric estimation in a homogeneous semi-Markov model. The waiting times distributions can be chosen to be the exponential, the Weibull, or the exponentiated Weibull distribution. Maximum likelihood estimations of both, hazard rates of the semi-Markov process and hazard rates of sojourn times can be deduced. Moreover, the effects of covariates on the process evolution can be studied using a semi-parametric Cox model for the distributions of sojourn times. The number of states, the possible transitions between them and the number of covariates affecting each transition can be chosen in order to fit sparse models adapted to a specific application. Package **SemiMarkov** is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=SemiMarkov`.

The rest of this paper is organized as follows. Section 2 describes the multi-state semi-Markov models and the parametric maximum likelihood estimation used in the **SemiMarkov** package. Section 3 describes the **SemiMarkov** package whereas the Section 4 illustrates the different functions included in the package through an example on severe asthma. Conclusions and possible future extensions of this R package are discussed in Section 5.

2. Homogeneous semi-Markov model framework

2.1. Homogeneous semi-Markov process

Let us consider a Markov renewal process $(J_n, T_n)_{n\in\mathbb{N}}$ where $0 = T_0 < T_1 < \ldots < T_n < \infty$ are the successive times of entry to states $J_0, J_1, \ldots, J_n$ where $J_n \neq J_{n+1}$ for all $n \in \mathbb{N}$. The sequence $(J_n)_{n\in\mathbb{N}}$ is an embedded homogeneous Markov chain taking values in a discrete finite state space E with transition probabilities $p_{hj} = \mathsf{P}(J_{n+1} = j|J_n = h)$, $n \in \mathbb{N}$. Let $S_n = T_n - T_{n-1}$ be the inter-arrival time for all $n \in \mathbb{N}^*$, $d \in \mathbb{R}_+$ and $(h, j) \in E \times E$, the Markov renewal kernel $Q_{hj}(d)$ satisfies

$$\begin{aligned} Q_{hj}(d) &= \mathsf{P}(J_{n+1} = j, S_{n+1} \leq d|J_0, \ldots, J_n = h, S_1, \ldots, S_n) \\ &= \mathsf{P}(J_{n+1} = j, S_{n+1} \leq d|J_n = h). \end{aligned} \tag{1}$$

Let $N(t) = \sup\{n \in \mathbb{N} : T_n \leq t, t \in \mathbb{R}_+\}$ be the counting process which counts the total number of observed transitions during the time interval $[0, t]$. The process $J_{N(t)}$, which represents the state of the process at time t, defines a homogeneous semi-Markov process.

The probability distribution function of sojourn times is related to the semi-Markov kernel through the transition probabilities of the embedded Markov chain,

$$F_{hj}(d) = \mathsf{P}(S_{n+1} \leq d|J_n = h, J_{n+1} = j) = \frac{Q_{hj}(d)}{p_{hj}}. \tag{2}$$

Let us suppose that the survival function $G_{hj}(.)$, the density function $f_{hj}(.)$ and the hazard rate $\alpha_{hj}(.)$ associated to this probability distribution can be defined. The survival function of sojourn time in state h is defined by $G_h(d) = 1-\mathsf{P}(S_{n+1} \leq d|J_n = h) = \sum_{j\in E} p_{hj}(1-F_{hj}(d))$. The hazard rate of the semi-Markov process corresponds to the probability of transition towards state j between time d and $d+\Delta d$, given that the process is in state h for a duration d

$$\lambda_{hj}(d) = \lim_{\Delta d\to 0} \frac{\mathsf{P}(J_{n+1} = j, d < S_{n+1} \leq d + \Delta d|J_n = h, S_{n+1} > d)}{\Delta d}. \tag{3}$$

The hazard of the semi-Markov process is related to the hazard rate of the sojourn time, the survival functions of the sojourn times and the transition probabilities of the Markov chain by the following relation

$$\begin{aligned} \lambda_{hj}(d) &= \frac{p_{hj}G_{hj}(d)\alpha_{hj}(d)}{G_h(d)}, \qquad h \neq j, \\ \lambda_{hh}(d) &= -\sum_{j\neq h} \lambda_{hj}(d). \end{aligned} \tag{4}$$

2.2. Sojourn times distribution

Let us assume that distributions of sojourn times belong to a parametric family. The simplest model is obtained using the exponential distribution $\mathcal{E}(\sigma_{hj})$, for which the hazard rate is constant over time (corresponding to the Markov case) and is related to a single positive parameter σ,

$$\alpha_{hj}(d) = \frac{1}{\sigma_{hj}}, \forall d \geq 0. \tag{5}$$

The Weibull distribution (Weibull 1951), which generalizes the exponential one, is often used in practical applications. Indeed, the Weibull distribution with two parameters $\mathcal{W}(\sigma_{hj}, \nu_{hj})$ is well adapted to deal with various shapes of monotone hazards,

$$\alpha_{hj}(d) = \frac{\nu_{hj}}{\sigma_{hj}} \left(\frac{d}{\sigma_{hj}} \right)^{\nu_{hj}-1}, \tag{6}$$

where $\sigma_{hj} > 0$ is a scale parameter and $\nu_{hj} > 0$ is a shape parameter. The exponentiated Weibull distribution $\mathcal{EW}(\sigma_{hj}, \nu_{hj}, \theta_{hj})$ (Mudholkar and Srivastava 1993) with an additional shape parameter $\theta_{hj} > 0$ is very useful to fit $\cup$ and $\cap$ shapes of hazard rates

$$\alpha_{hj}(d) = \frac{\theta_{hj} \frac{\nu_{hj}}{\sigma_{hj}} \left(\frac{d}{\sigma_{hj}} \right)^{\nu_{hj}-1} \exp\left(- \left(\frac{d}{\sigma_{hj}} \right)^{\nu_{hj}} \right) \left[1 - \exp\left(- \left(\frac{d}{\sigma_{hj}} \right)^{\nu_{hj}} \right) \right]^{\theta_{hj}-1}}{\left(1 - \left[1 - \exp\left(- \left(\frac{d}{\sigma_{hj}} \right)^{\nu_{hj}} \right) \right]^{\theta_{hj}} \right)}. \tag{7}$$

These three distributions which allow to fit various shapes of the hazard ratio are nested: a $\mathcal{EW}(\sigma_{hj}, 1, 1)$ is equivalent to a $\mathcal{W}(\sigma_{hj}, 1)$ which is equivalent to a $\mathcal{E}(\sigma_{hj})$. The Wald test can be used to test each parameter and evaluate the relevance of a given distribution.

2.3. Parametric maximum likelihood estimation

In a parametric framework, distributions of sojourn times are supposed to belong to a class of parametric functions. For each transition, the distribution (which depends on a finite number of parameters) can be specified using either the hazard rate $\alpha_{hj}(\cdot)$, the density $f_{hj}(\cdot)$ or the cumulative distribution function $F_{hj}(\cdot)$.

The likelihood function associated to a single semi-Markov process can be written as

$$L = \left[\prod_{n=1}^{N} p_{J_{n-1}J_n} f_{J_{n-1}J_n}(S_n) \right] \times \left[G_{J_N}(U) \right]^{\delta}, \tag{8}$$

where N is the total number of observed transitions between two different states, U denotes the duration between the time of the last observation and T_N the time of the end of the study. The indicator δ is equal to 1 if the last sojourn time U is right-censored by the end of the study. Indeed, the last duration and the last arrival state are unknown unless the process entered an absorbing state ($\delta = 0$). When an observation is right-censored ($\delta = 1$), the survival function of the sojourn times is taken into account. The first part of Equation 8 involves the density function and the probabilities of the Markov chain; it corresponds to the contribution of the observed transitions.

Consider that each individual $i = 1, \ldots, k$, is associated to a semi-Markov process $(J_{N_i(t)}, t \geq 0)$ with $N_i(t) = \sup\{n \in \mathbb{N} : T_n^i \leq t,\ t \in \mathbb{R}_+\}$. Equation 8 can be used to compute each individual contribution to the likelihood L_i. The full likelihood L is the product of all individual likelihood contributions L_i.

2.4. Cox proportional model

The influence of covariates on the sojourn times distributions can be studied using a Cox proportional regression model (Cox 1972). Let Z_{hj} be a vector of explanatory variables and

β_{hj} a vector of regression parameters associated to the transition from state h to state j. The hazard rate is defined by

$$\alpha_{hj}(d|Z_{hj}) = \alpha_{hj0}(d)\exp(\beta_{hj}^{\top}Z_{hj}), \quad d \geq 0,\ h, j \in E,\ h \neq j, \tag{9}$$

where $\alpha_0(d)$ denotes the baseline hazard defined in Section 2.2.

In this model, the regression coefficients can be interpreted in terms of relative risk. As in the Cox model, time-dependent covariates can be considered assuming that the value of the covariate is constant between two consecutive events. Let us mention that the previous notation allows to consider different sets of covariates for each transition. It is then possible to consider sparse models with only significant regression parameters.

3. The SemiMarkov R package

3.1. Package description

The **SemiMarkov** package was developed to analyze longitudinal data using multi-state semi-Markov models. The main function `semiMarkov` of the package computes the parametric maximum likelihood estimation in the homogeneous semi-Markov model introduced in Section 2.

Format of data

A data set `asthma` is included in the **SemiMarkov** package. This cohort study (longitudinal data) of severe asthmatic patients can be analyzed using multi-state semi-Markov models. The data frame to be used in the function `semiMarkov` must be similar to the `asthma` data: a table in long format (one row per transition and possibly several rows by individual) that must contain the following information

1. `id`: The individual identification number.
2. `state.h`: State left by the process.
3. `state.j`: State entered by the process.
4. `time`: Sojourn time in `state.h`.

The rows must be grouped by individuals and ordered chronologically within groups. By definition of a semi-Markov model the waiting times must be known. Therefore, transitions between the same states are not possible. If such transitions are observed, the row must be combined with the next transition to obtain a transition from state h to state j with $h \neq j$. The last sojourn time of a semi-Markov process is observed only when the process enters an absorbing state. In other cases, the final state and the last sojourn time is unknown due to the right-censoring process. In such case, it is only known that the censored sojourn time is greater than the last observed sojourn time (in practice, the last observed sojourn time is deduced from the date of the end of the study). A censored transition can be specified by a transition from h to h (so that such transitions are distinct from the rest of the transitions). One can also identify the unknown arrival state using the argument `cens`. The data set may

also include additional explanatory variables (for instance, some individual's characteristics). The values of these covariates must be given for each individual and for each transition in order to take fixed or time-dependent covariates into account (one value for each row of the data frame `data`).

Functions description

Following is a brief description of the package functions.

- `table.state`: Computes a frequency table counting the number of observed transitions in the data set.
- `param.init`: Defines default or specified initial values of the parameters.
- `semiMarkov`: Computes the parametric maximum likelihood estimation of multi-state semi-Markov models.
- `hazard`: For any object of classes '`semiMarkov`' and '`param.init`', the function computes the values of the hazard rate of sojourn times or the values of the hazard rate of the semi-Markov process for a given vector of times.
- `summary` and `print`: Summary and printing methods for objects of classes '`semiMarkov`' and '`hazard`'.
- `plot`: Plot method for objects of class '`hazard`'.

Sojourn times distribution

The parametric estimation in homogeneous semi-Markov models is based on the specification of the sojourn times distribution. The following distributions are available in the package **SemiMarkov**: exponential (`"E"`, `"Exp"` or `"Exponential"`), Weibull (`"W"` or `"Weibull"`) and exponentiated Weibull (`"EW"`, `"EWeibull"` or `"Exponentiated Weibull"`). If the logical value `TRUE` is given then the default is the Weibull distribution. These distributions are nested when the appropriate parameters are equal to 1 (see Section 2). The estimations of the distribution parameters are given with standard deviations and p values of the Wald test ($H_0 : \Theta_{hj} = 1$ is the default null hypothesis). One can then evaluate, for instance, the relevance of the exponentiated Weibull distribution in comparison to the Weibull or the exponential distribution.

Multi-state model definition

The multi-state approach requires to define the states of the process and to specify the structure of the model (the number of states and the possible transitions between them). The function `table.state` returns a matrix which gives the number of observed transitions in the data set. This function can help to define the argument `mtrans` required in the `semiMarkov` and the `hazard` functions. The square matrix `mtrans` includes information on possible transitions and on the distributions of waiting times. The element hj of the matrix `mtrans` is either a logical value `FALSE` (when the transition from h to j is not possible) or a character representing the sojourn time distribution. According to semi-Markov models, the diagonal

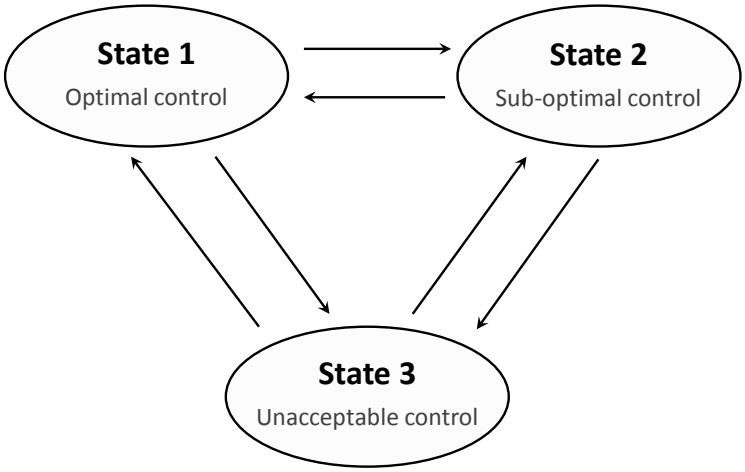

Figure 1: The three states model used for asthma control evolution.

elements of `mtrans` are all equal to `FALSE`. Note that only the transitions specified in `mtrans` will be considered in the analysis. In case of the three-state model described in Figure 1 where the sojourn times associated to each transition are Weibull distributed, the matrix `mtrans` will be defined as follows

```
R> mtrans
```

```
     [,1]    [,2]    [,3]
[1,] "FALSE" "W"     "W"
[2,] "W"     "FALSE" "W"
[3,] "W"     "W"     "FALSE"
```

The argument `states` is a character vector used to define the names of states, possible values are those included in the data's columns `state.h` and `state.j`.

Covariates

The effect of covariates on the process evolution can be investigated considering a Cox proportional hazards model for the hazard rates of waiting times. A set of covariates can be specified using the argument `cov`. The argument `cov_tra` is used to indicate which covariates affect which transitions: `cov_tra` is a list of vectors where the kth vector provides the transitions affected by the kth covariate. The elements of these vectors can only consist of transitions specified in the argument `mtrans`. For instance, let us consider a three states model where all transitions between states are possible (Figure 1) and let us suppose that a covariate named `Cov1` affects all transitions leaving state 1 whereas a second covariate named `Cov2` affects the transitions leaving state 2: in this case, the arguments to be passed in the function must be respectively `cov = data.frame(Cov1, Cov2)` and `cov_tra = list(c("12", "13"), c("21", "23"))`. The interpretation of the regression coefficients in terms of relative risks (as in the Cox model) can help to quantify the effect of covariates and to understand the process evolution. For each estimation of regression coefficients, standard deviation and p value of the Wald test ($H_0 : \beta_{hj} = 0$ is the default null hypothesis) are given.

Initial values

The optimization procedure used in the maximum likelihood estimation requires definition of initial values of the parameters: the distribution parameters, the transition probabilities and the regression coefficients associated to the covariates. Default values are equal to 1 for the distribution parameters, and 0 for the regression coefficients. The initial transition probabilities are calculated by simple proportions: the number of observed transitions from state h to state j divided by the total number of observed transitions from state h. The function `param.init` can be used to define specific initial values of the parameters. An object of class '`param.init`' can then be given as argument in the `semiMarkov` and `hazard` functions. The total number of parameters depends on: the number of states, the possible transitions, the chosen distributions and the covariates.

3.2. Parametric maximum likelihood estimation

The `semiMarkov` *function*

The main function `semiMarkov` estimates the parameters of a multi-state homogeneous semi-Markov model using parametric maximum likelihood estimation. Several R packages are needed to run the function. The package **numDeriv** (Gilbert and Varadhan 2015) that allows to approximate the Hessian matrix of second derivatives for the estimated parameters. The package **MASS** (Venables and Ripley 2002) is used to obtain the inverse of the Hessian matrix. The maximization of the likelihood function is performed using non-linear optimization based on the augmented Lagrange multiplier method implemented in the function `solnp` from the R package **Rsolnp** (Ghalanos and Theussl 2014). Indeed, we have to face an optimization with constraints since sums of the probabilities associated to transitions leaving the same states are all equal to 1 (the sums in rows of the transition matrix).

The following arguments are used in the function `semiMarkov`: arguments related to the data (`data`, `cov`), arguments related to the model (`states`, `mtrans`, `cov_tra`, `cens`) and initial values (`dist_init`, `proba_init`, `coef_init`). Default values are defined for the distributions of waiting times and for the initial values. The function `semiMarkov` returns an object of class '`semiMarkov`' which recalls the chosen model, gives information on the optimization method and provides the parameter estimates together with their standard deviations. For each regression parameter β and distribution parameter σ (or ν or θ), the function `SemiMarkov` also provides the Wald test statistic and p value associated to a given null hypothesis which can be specified using argument `Wald_par`. The default null hypothesis for regression coefficients is the absence of association ($H_0 : \beta = 0$) whereas the default null hypothesis for distribution parameters is $H_0 : \sigma = 1$. The Wald test for the transition probabilities is less useful and is not performed.

Some arguments related to the optimization procedure can also be specified in the function `semiMarkov` and passed further to the function `solnp`. Indeed, the function `solnp` allows to define constraints on the model's parameters. The arguments `ineqLB` and `ineqUB` can be used in the `semiMarkov` function to impose respectively lower and upper bounds on the parameter estimates. The argument `eqfun` can be used to define constraints of type $par_1 = a \times par_2$ where a is a constant and par_1 and par_2 are two parameters. Note that it is only possible to specify constraints between the same type of parameters (distribution parameters, transition probabilities or regression coefficients).

The hazard function

The hazard rate of sojourn time and the hazard rate of the semi-Markov process can be deduced from the parameters and the distributions of sojourn times using Equation 9 and Equation 3, respectively. The function `hazard` computes vectors of hazard rates values using either the estimates included in an object of class '`semiMarkov`' or the specific values defined by an object of class '`param.init`'. The argument `type` is used to choose the type of hazard rate: `alpha` for the hazard rates of waiting times and `lambda` for the hazard rates of the semi-Markov process.

The `hazard` function returns the values of the hazard rates associated to a vector of times. By default, the hazard rates are calculated for a vector of ordered times of length 1000 where the starting value is equal to 0 and the ending value is determined by the longest sojourn times. The length of the vector, its starting and ending values can be specified by the user. One can also enter a whole vector of times, for instance, the different values of the sojourn times observed in the data. If covariates are used in the model, the hazard rates can be obtained for given values of the covariates using the argument `cov`: for time-fixed covariates a single value is needed whereas a vector of values is required for time-dependent covariates. By default, all covariate values are set equal to 1. Note that the function `hazard` does not require to specify the model or the distributions. Indeed, this information is already included in the objects of class '`semiMarkov`' or '`param.init`'.

3.3. Showing results

An object of class '`semiMarkov`' contains the data description, the considered model and the results of the maximum likelihood estimation that may be displayed using the `summary` or `print` method for '`semiMarkov`' objects. The `summary` and `print` methods for '`hazard`' objects provide the type of hazard rates, the vector of times and the associated values of hazard rates. An object of the class '`hazard`' can be plotted using the corresponding `plot` method. For each transition, the function generates a plot representing one or more (up to ten) hazard rates.

4. Application to asthma control data

A follow-up study of severe asthmatic patients was conducted in France between 1997 and 2001 by ARIA (Association pour la Recherche en Intelligence Artificielle). Adult asthmatics were prospectively enrolled over a 4-year period by a number of French chest physicians. The data reflects the real follow-up of patients consulting at varied times according to their perceived needs. At each visit, several covariates were recorded and asthma was evaluated using the concept of control scores (Saint-Pierre, Bourdin, Chanez, Daures, and Godard 2006). The control scores can be used to define the subject's state at each consultation. The considered model to study the evolution of asthma consists of three transient states (Figure 1): the optimal control (State 1), the sub-optimal control (State 2) and the unacceptable control (State 3).

A random selection of 371 patients with at least two visits (data `asthma`) is included in the package **SemiMarkov**. A total of 557 transitions between states are observed and no deaths are reported. Together with the control scores at each time, three covariates are included

in the data: severity (disease severity: coded 1 if severe, 0 if mild-moderate asthma), BMI (body mass index: 1 if BMI $\geq$ 25, 0 if BMI $<$ 25) and sex (1 if men, 0 if women). The data frame `asthma` contains one row per transition. The rows corresponding to the same subject are grouped and ordered chronologically. The columns of the `asthma` data are: the patient identification number (`id`), the state left by the process (`state.h`), the arrival state (`state.j`), the sojourn time in state `state.h` (`time`) and binary covariates (`Severity`, `BMI`, `Sex`). Note that the variable `BMI` is a time-dependent covariate.

```
R> library("SemiMarkov")
R> data("asthma", package = "SemiMarkov")
R> head(asthma)
```

```
  id state.h state.j       time Severity BMI Sex
1  2       3       2 0.15331964        1   1   0
2  2       2       2 4.12320329        1   1   0
3  3       3       1 0.09582478        1   1   1
4  3       1       3 0.22997947        1   1   1
5  3       3       1 0.26557153        1   1   1
6  3       1       1 5.40725530        1   1   1
```

There are no absorbing states in the considered model (Figure 1). The last sojourn time is then right-censored. Its value is the time between the last visit and the date of the end of the study. A censored observation is identified by a transition into the same state. In such case, the value of `state.h` is equal to the value of `state.j` and the value of `time` is the censored sojourn time.

```
R> table.state(asthma)
```

```
$table.state
    1   2  3
1 152  95 44
2 112 116 71
3 115 120 103
$Ncens
[1] 371
```

In a primary analysis, the data are stratified according to the values of the covariates. The effect of covariates and the proportional hazard assumption can be evaluated by representing the hazard rates in each stratum. In a second step, a proportional model can be considered to study the effect of covariates. For instance, one can consider a model with BMI as covariate and the Weibull distribution for the waiting times.

```
R> states <- c("1", "2", "3")
R> mtrans <- matrix(FALSE, nrow = 3, ncol = 3)
R> mtrans[1, 2:3] <- c("W", "W")
R> mtrans[2, c(1, 3)] <- c("W", "W")
R> mtrans[3, c(1, 2)] <- c("W", "W")
```

```
R> BMI <- as.data.frame(asthma$BMI)
R> fit <- semiMarkov(data = asthma, states = states, mtrans = mtrans,
+    cov = BMI)
```

The `semiMarkov` function provides estimations of parameters of the waiting times distributions, the standard deviations, the confidence intervals and the Wald test statistics (H_0 : $\theta_{hj} = 1$). One can observe that the coefficient ν_{23} associated to the transition from state 2 to state 3 is not significantly different from 1. The exponential distribution can then be used instead of the Weibull distribution for this transition.

```
R> fit$table.dist
```

```
$Sigma
  Type Index Transition Sigma   SD Lower_CI Upper_CI Wald_H0 Wald_test
1 dist     1     1 -> 2 9.384 2.42     4.64    14.13    1.00     12.01
2 dist     2     1 -> 3 0.418 0.08     0.26     0.58    1.00     51.54
3 dist     3     2 -> 1 5.014 1.25     2.57     7.46    1.00     10.36
4 dist     4     2 -> 3 0.714 0.12     0.49     0.94    1.00      6.06
5 dist     5     3 -> 1 2.233 0.53     1.20     3.26    1.00      5.51
6 dist     6     3 -> 2 0.498 0.08     0.34     0.65    1.00     41.07
  p_value
1  0.0005
2 <0.0001
3  0.0013
4  0.0138
5  0.0189
6 <0.0001

$Nu
  Type Index Transition    Nu   SD Lower_CI Upper_CI Wald_H0 Wald_test
1 dist     7     1 -> 2 0.531 0.05     0.44     0.63    1.00     95.79
2 dist     8     1 -> 3  1.18 0.14     0.90     1.46    1.00      1.65
3 dist     9     2 -> 1  0.51 0.04     0.43     0.59    1.00    142.11
4 dist    10     2 -> 3 1.048 0.10     0.86     1.24    1.00      0.25
5 dist    11     3 -> 1 0.499 0.04     0.42     0.58    1.00    161.12
6 dist    12     3 -> 2 0.931 0.06     0.81     1.06    1.00      1.14
  p_value
1 <0.0001
2  0.1990
3 <0.0001
4  0.6171
5 <0.0001
6  0.2857
```

The regression coefficients associated with BMI can be analyzed using the Wald test statistics ($H_0 : \beta_{hj} = 0$). For instance, the estimate of the coefficient associated to the transition from state 3 to state 1 is significantly different from 0 ($\beta = -0.447, p = 0.028$). It means that a

BMI ≥ 25 decreases the risk of leaving the unacceptable state to enter the optimal control state.

```
R> fit$table.coef
```

```
> fit$table.coef
  Type Index Transition Covariates  Estimation   SD Lower_CI Upper_CI Wald_H0
1 coef     1     1 -> 2      Beta1 -0.27808218 0.22    -0.72     0.16    0.00
2 coef     2     1 -> 3      Beta1 -0.87827431 0.35    -1.57    -0.19    0.00
3 coef     3     2 -> 1      Beta1  0.03216304 0.19    -0.35     0.41    0.00
4 coef     4     2 -> 3      Beta1 -0.11151373 0.27    -0.64     0.41    0.00
5 coef     5     3 -> 1      Beta1 -0.61127841 0.20    -1.00    -0.22    0.00
6 coef     6     3 -> 2      Beta1 -0.23912937 0.21    -0.65     0.17    0.00
  Wald_test p_value
1      1.55  0.2131
2      6.27  0.0123
3      0.03  0.8625
4      0.17  0.6801
5      9.43  0.0021
6      1.32  0.2506
```

The effect of BMI on the hazards of waiting times and on the hazards of the semi-Markov process can also be evaluated using the `hazard` and `plot` functions (Figure 2).

```
R> plot(hazard(fit, cov = 0), hazard(fit, cov = 1), transitions = "13")
R> plot(hazard(fit, cov = 0, type = "lambda"), hazard(fit, cov = 1,
+    type = "lambda"), transitions = "13", legend.pos = c(3.75, 0.119),
+    cex = 0.8)
```

Finally, multivariate models can be considered. However, the number of parameters can quickly be too large compared to the size of the data set (due to the complex form of the waiting times distributions and to the number of covariates under study). The optimization method can then fail to reach convergence. It is therefore important to consider sparse models using adapted distributions and keeping only the regression coefficients significantly different from 1. One can also specify the null hypothesis of the Wald test using argument `Wald_par`. In the following example, the null hypothesis are the nullity of distribution parameters and the regression coefficients are equal to -1.

```
R> SEV <- as.data.frame(asthma$Severity)
R> fit2 <- semiMarkov(data = asthma, cov = data.frame(BMI, SEV),
+    states = states, mtrans = mtrans, cov_tra = list(c("13", "31"),
+    "23"), Wald_par = c(rep(0, 12), rep(-1, 3)))
```

Constraints on the lower and upper bounds for parameters can be specified. For instance, the estimation of the regression coefficient associated to transition from state 1 to state 2 can belong to $[-0.2, 0.2]$ whereas the regression coefficient associated to transition from state 2 to state 3 can belong to $[-0.1, 0.1]$.

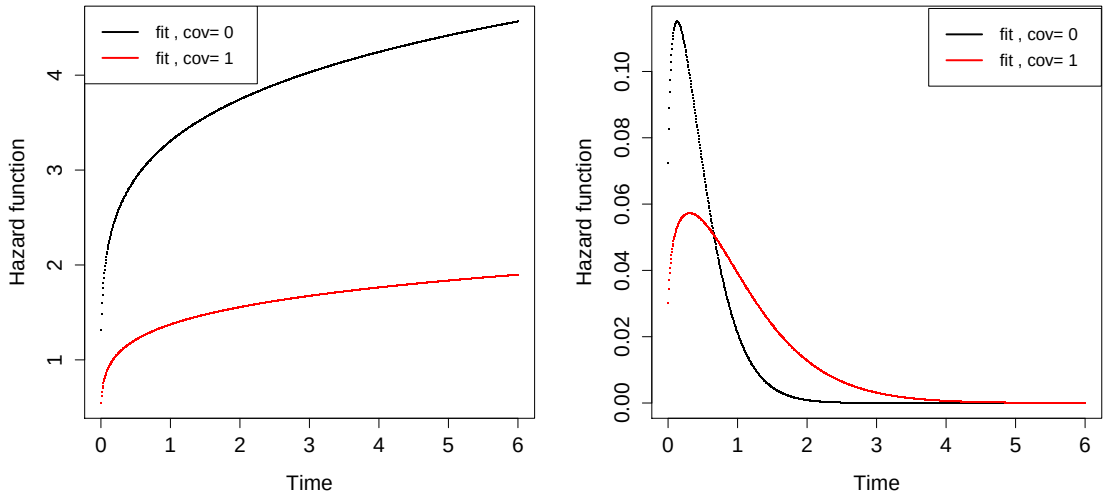

Figure 2: The hazard rate of waiting time (left) and the hazard rate of the semi-Markov process (right) for the transition 13 for BMI = 0 (black line) and for BMI = 1 (red line).

```
R> fit3 <- semiMarkov(data = asthma, cov = BMI, states = states, mtrans =
+    mtrans, ineqLB = list(c("coef", 1, -0.2), c("coef", 4, -0.1)),
+    ineqUB = list(c("coef", 1, 0.2), c("coef", 4, 0.1)))
```

The definition of arguments `ineqLB` and `ineqUB` requires three elements: the type of parameters (`"dist"`, `"proba"` or `"coef"`), the index of the parameter (can be identified from the `semiMarkov` output) and the lower (or upper) bound. One can also impose equality constraints on the parameters. It is of particular interest to suppose that two (or several) distribution parameters or regression coefficients are equal. In the following example, the second distribution parameters associated to transition from state 1 to state 2 and and from state 2 to state 1 are equal to each other.

```
R> fit4 <- semiMarkov(data = asthma, cov = BMI, states = states, mtrans =
+    mtrans, eqfun = list(c("dist", 7, 9, 1)))
```

The definition of argument `eqfun` requires four elements: the type of parameters (`"dist"`,`"proba"` or `"coef"`), the index of the first parameter (can be identified from the `semiMarkov` output), the index of the second parameter and a constant a such that $par_1 = a \times par_2$. In the following example it is supposed that the regression coefficients associated to transition from state 2 to state 1, from state 2 to state 3 and from state 3 to state 2 are equal to each other.

```
R> fit5 <- semiMarkov(data = asthma, cov = BMI, states = states, mtrans =
+    mtrans, eqfun = list(c("coef", 3, 4, 1), c("coef", 3, 6, 1)))
```

Note that additional constraints on transition probabilities must be chosen with caution since these parameters are already subject to constraints induced by the model definition. Indeed, the probability must obviously belong to $[0, 1]$ and the sum of the probabilities in the same row of the transition matrix must be equal to one (the last probability of a given row is not estimated but deduced from the others). Therefore, no additional constraints related to these probabilities are permitted.

5. Discussion

Semi-Markov multi-state models are proven to be very useful in various applications. They are extensions of Markov models in which the evolution of a process is independent from time spent in a state between two consecutive events. Such an assumption is too stringent in some applications. In such case, semi-Markov models are of great interest for modeling the sojourn (waiting) times distributions. However, the implementation of such approach is complex and there are barely any packages or macros to adjust such models. The **SemiMarkov** package allows to fit parametric homogeneous semi-Markov models by maximizing the likelihood. The choice of waiting times distributions, in particular the exponentiated Weibull distribution, allows to fit various shapes of hazard rates functions. An advantage of the parametric approach is the possibility to study the effects of covariates via a proportional hazard model. In order to obtain sparse models adapted to the process of interest, the user can choose the number of covariates and the distributions of waiting times for each transition. Some extensions of the **SemiMarkov** package could be of interest. For instance, the package could be updated to deal with more waiting times distributions. The methodology can be adapted to include random effects in order to deal with the correlation between subjects. Interval censored data could also be analyzed using a penalized likelihood approach (Foucher, Giral, Soulillou, and Daures 2010) or using an estimation method with piecewise constant hazard rates (Kapetanakis, Matthews, and Hout 2012). The optimization step is a crucial point and needs to be investigated. Indeed, the multi-state approach is often limited by the number of parameters with several covariates. The adaptation of methods dealing with high dimensional data to the multi-state model framework is of high interest as well.

References

Aguirre-Hernàndez R, Farewell VT (2002). "A Pearson-Type Goodness-of-Fit Test for Stationary and Time-Continuous Markov Regression Models." *Statistics in Medicine*, **21**(13), 1899–1911.

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. URL http://www.jstatsoft.org/v38/i04/.

Cox DR (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistics Society B*, **34**(2), 187–220.

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/.

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. URL http://www.jstatsoft.org/v50/i14/.

Foucher Y, Giral M, Soulillou JP, Daures JP (2010). "A Flexible Semi-Markov Model for Interval-Censored Data and Goodness-of-Fit Testing." *Statistical Methods in Medical Research*, **19**(2), 127–145.

Foucher Y, Mathieu E, Saint-Pierre P, Durand JF, Daurès JP (2005). "A Semi-Markov Model Based on Generalized Weibull Distribution with an Illustration for HIV Disease." *Biometrical Journal*, **47**(6), 825–833.

Ghalanos A, Theussl S (2014). ***Rsolnp**: General Non-Linear Optimization*. R package version 1.15, URL http://CRAN.R-project.org/package=Rsolnp.

Gilbert P, Varadhan R (2015). ***numDeriv**: Accurate Numerical Derivatives*. R package version 2014.2-1, URL http://CRAN.R-project.org/package=numDeriv.

Gill RD (1980). "Nonparametric Estimation Based on Censored Observations of Markov Renewal Process." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **53**(1), 97–116.

Huszti E, Abrahamowicz M, Alioum A, Binquet C, Quantin C (2012). "Relative Survival Multistate Markov Model." *Statistics in Medicine*, **31**(3), 269–286.

Jackson CH (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/.

Kapetanakis V, Matthews FE, Hout A (2012). "A Semi-Markov Model for Stroke with Piecewise-Constant Hazards in the Presence of Left, Right and Interval-Censoring." *Statistics in Medicine*, **32**(4), 697–713.

Limnios N, Oprisan G (2001). *Semi-Markov Processes and Reliability*. Springer-Verlag.

Listwon-Krol A, Saint-Pierre P (2015). ***SemiMarkov**: Multi-States Semi-Markov Models*. R package version 1.4.2, URL http://CRAN.R-project.org/package=SemiMarkov.

Maechler M (2015). ***VLMC**: Variable Length Markov Chains*. R package version 1.4-1, URL http://CRAN.R-project.org/package=VLMC.

Meira-Machado L, Pardiñas JR (2011). "**p3state.msm**: Analyzing Survival Data from an Illness-Death Model." *Journal of Statistical Software*, **38**(3), 1–18. URL http://www.jstatsoft.org/v38/i03/.

Monteiro A, Smirnov GV, Lucas A (2006). "Non-Parametric Estimation for Nonhomogeneous Semi-Markov Processes: An Application to Credit Risk." *Discussion Paper TI 2006-024/2*, Tinbergen Institute. URL http://papers.tinbergen.nl/06024.pdf.

Mudholkar GS, Srivastava DK (1993). "Exponentiated Weibull Family for Analyzing Bathtub Failure-Rate Data." *IEEE Transactions on Reliability*, **42**(2), 299–302.

O'Connell J, Højsgaard S (2011). "Hidden Semi Markov Models for Multiple Observation Sequences: The **mhsmm** Package for R." *Journal of Statistical Software*, **39**(4), 1–22. URL http://www.jstatsoft.org/v39/i04/.

Ouhbi B, Limnios N (1999). "Nonparametric Estimation for Semi-Markov Processes Based on its Hazard Rate Functions." *Statistical Inference for Stochastic Processes*, **2**(2), 151–173.

Pérez-Ocón R, Ruiz-Castro JE, Gàmiz-Pérez ML (1999). "Semi-Markov Models for Lifetime Data Analysis." In *Semi-Markov Models and Applications*, pp. 229–238. Springer-Verlag.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Saint-Pierre P, Bourdin A, Chanez P, Daures JP, Godard P (2006). "Are Overweight Asthmatics More Difficult to Control?" *Allergy*, **61**(1), 79–84.

Saint-Pierre P, Combescure C, Daurès JP, Godard P (2003). "The Analysis of Asthma Control under a Markov Assumption with Use of Covariates." *Statistics in Medicine*, **22**(24), 3755–3770.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Weibull W (1951). "A Statistical Distribution Function of Wide Applicability." *Journal of Applied Mechanics*, **18**, 292–297.

Affiliation:

Agnieszka Król
INSERM U897 (Biostatistics Team)
Université de Bordeaux
33076 Bordeaux, France
E-mail: agnieszka.krol@isped.u-bordeaux2.fr

Philippe Saint-Pierre
Laboratoire de Statistique Théorique et Appliquée
Université Pierre et Marie Curie
75252 Paris, France
E-mail: philippe.saint_pierre@upmc.fr

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 66, Issue 6 *Submitted:* 2013-05-14
August 2015 *Accepted:* 2014-10-21