The SmoothHazard package for R: Fitting regression models to interval-censored observations of illness-death models

Célia Touraine
University of Bordeaux[1], Thomas A. Gerds[2], and Pierre Joly[3]

[1]INSERM Epidemiology and Biostatistics Research Center,Bordeaux,France, E-mail: celia.touraine@isped.u-bordeaux2.fr
[2]Department of Biostatistics, University of Copenhagen, Denmark
[3]INSERM, Epidemiology and Biostatistics Research Center, Bordeaux, France

October 25, 2013

1 Introduction

The irreversible illness-death model is a multi-state model which has many applications in various areas of research, for example in the medical field. The model describes the transitions from an initial state (e.g., alive and disease-free) to an absorbing state (e.g., death) either directly or via an intermediate state (e.g., disease) (Figure 1). The transition intensities α_{01}, α_{02}, and α_{02} are positive functions of time which can also depend on covariates.

In some applications it happens for some or all subjects that the transition times from the initial state to the intermediate state are interval censored. This occurs for example when the status of the intermediate state can only be determined at a sequence of visit times. In this case, if a subject is diagnosed as diseased at one of the visit times, say R, then it is only known that the subject was last seen disease-free at the previous visit time, say L, and hence the time of the onset of the disease is interval censored between L and R for this subject. Furthermore, both the process of visit times and the observation of the time of the transition into the absorbing state are usually right censored, i.e., limited to the individual follow-up period of the subjects. This yields a rather complex general observational pattern, because for a subject who died without being diagnosed as diseased at earlier visit times, it may or it may not be possible to determine retrospectively if and when the subject became diseased between the last visit time and the time of death.

The **SmoothHazard** package provides estimates of the baseline transition intensities and of covariate effects when the data fall into one of the 6 cases that

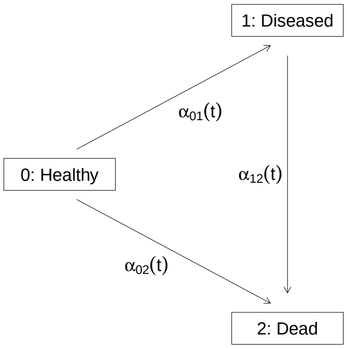


Figure 1: The irreversible illness-death model

are displayed in Figure 2. Thus, the case of left-truncated event times (delayed entry) is covered, as well as the case where for some subjects the transition time into the intermediate state is observed exactly and for others it is interval censored. Finally, the special case is covered where for some or all subjects no intermediate information is available about the disease status such that it is only known whether or not the subjects became diseased between the start and the end of follow-up. The latter occurs in Figure 2 when $E = L$ and $R = \min(T, C)$ in cases 2. or 4.

To estimate covariate effects on the three transition intensities, implemented are regression models which assume proportional transition intensities and a non-homogeneous Markov process. The user chooses between a fully parametric model where each of the baseline intensities is described by the parameters of a Weibull distribution and a semi-parametric model where the baseline intensities are left unspecified and approximated by M-splines. For the parametric model, the regression coefficients and Weibull parameters are estimated by maximising the likelihood, for the semi-parametric model, the coefficients of the M-splines and the regression coefficients are estimated by maximising a penalized likelihood.

The package **SmoothHazard** then allows to do predictions of transition probabilities, cumulative probabilities of event and life expectancies for a given set of covariates, based on estimated baseline transition intensities and on estimated covariates effects.

If the exact transition times are observed, standard procedures like those implemented, e.g. in the packages **survival**, **rms**, **etm**, **mstate** can be used to estimate transition intensities, regression coefficients and functionals thereof [see 4, 2]. In particular, the regression coefficients can be estimated using Cox partial likelihood [3] without the need to model the baseline intensities. However, when transition times to the intermediate event are interval censored, it is generally not possible to arrive at consistent estimates with the software provided by the packages listed above. Indeed, the approach to handle subjects who died with unknown disease status, consists in artificially ending their follow-up at the last time they were seen without disease and subsequently treat them as right-censored. However, this ap-

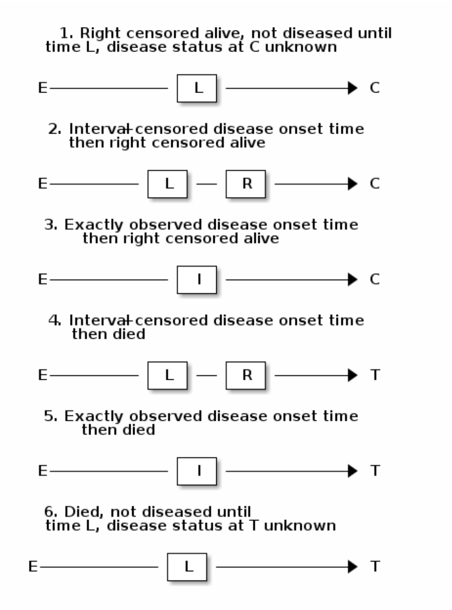

1. Right censored alive, not diseased until time L, disease status at C unknown
E
L
C
2. Interval-censored disease onset time then right censored alive
E
L
R
C
3. Exactly observed disease onset time then right censored alive
E
I
C
4. Interval-censored disease onset time then died
E
L
R
T
5. Exactly observed disease onset time then died
E
I
T
6. Died, not diseased until time L, disease status at T unknown
E
L
T

Figure 2: Observational patterns that are recognized by **SmoothHazard**. The letters I and T denote the transition times into the intermediate and absorbing state, respectively. The letters E and C denote the start and end of follow-up, respectively, and the letters L and R the visit times between which the transition into the intermediate happened.

proach can lead to a systematic bias in the estimates of transition intensities and of regression coefficients [6, 7]. The bias will be especially pronounced if the risk of death is higher for diseased subjects than the risk of death for disease-free subjects.

The **msm** package [5] allows to fit Markov multi-state models to panel data where the status of the subjects is known at a finite series of inspection times. As a special case the setting includes the illness-death model and it can be used with interval-censored disease times and exact death times. However, in this package the likelihood is calculated using the Kolmogorov differential equations that relate the transition probabilities and the transition intensities and to make this work a time-homogeneity assumption is made where all transition intensities are constant or piecewise-constant between two successive observation times.

1.0.1 Outline

The main functions of **SmoothHazard** is

- `idm` : for fitting illness-death regression models based on possibly interval-censored disease times and right censored times.

A fitted illness-death model as produced by idm can be used in the following functions to calculate predictions:

- predict.idm : for estimating transition probabilities and cumulative probabilities of event for a given set of covariates;
- lifexpect : for estimating life expectancies for a given set of covariates.

The R function idm is essentially an interface between the user and *F*ORTRAN programs which constitute the heart of the package **SmoothHazard**.

Section 2 presents the model and the likelihood. Section 3 presents the estimation methods. Section 4 briefly presents predictions that can be made in an illness-death model. Section 5 provides some examples illustrating **SmoothHazard**.

2 Model and likelihood

We consider an illness-death process $X = (X(t), t \geq 0)$ which takes values in $\{0, 1, 2\}$ (Figure 1). Subjects are initially disease-free ($X(0) = 0$) and may become diseased (transition $0 \rightarrow 1$) and die (transition $1 \rightarrow 2$), or die directly without disease (transition $0 \rightarrow 2$.) X is assumed to be a non-homogeneous Markov process which means that the future evolution of the process $\{X(t), t > s\}$ depends on the current time s and only on the current state $X(s)$. Thus, the distribution of X is fully characterized by the set of transition probabilities:

$$p_{hl}(s,t) = \mathrm{P}(X(t) = l | X(s) = h) \qquad hl \in \{01, 02, 12\}.$$

The transition probabilities are related to the instantaneous transition intensities α_{hl} shown in Figure 1 by the relation:

$$\alpha_{hl}(t) = \frac{p_{hl}(t, t + \Delta t)}{\Delta t}.$$

We introduce covariate effects separately for each transition through proportional transition intensities regression models which are a natural extension of the Cox proportional hazard model:

$$\alpha_{hl}(t|Z_{hli}) = \alpha_{0,hl}(t) \exp\{\beta_{hl}^T Z_{hli}\}; \qquad hl \in \{01, 02, 12\}. \tag{1}$$

Here $\alpha_{0,hl}$ are baseline transition intensities, Z_{hli} are covariate vectors for subject i and β_{hl} are vectors of regression parameters for transition $h \rightarrow l$.

In the situation where the time to disease and the time to death are not interval censored but either observed exactly or right censored, the regression coefficients can be estimated by the partial likelihood method without the need to specify and estimate the baseline hazard functions $\alpha_{0,hl}(t)$. For interval-censored transition times to the intermediate state, the situation is more complex. It turns out that we have to estimate all parameters simultaneously and that we need a model for the baseline transition intensity functions. This can be seen by inspecting the likelihood function.

For subject i, denote the conditional disease-free survival function by

$$S(t|Z_{01i}, Z_{02i}) = e^{-A_{01}(t|Z_{01i})-A_{02}(t|Z_{02i})}$$

where $A_{hl}(.|Z_{hli})$ is the conditional cumulative intensity function of transition $h \rightarrow l$:

$$A_{hl}(t|Z_{hli}) = \int_0^t \alpha_{hl}(u|Z_{hli})du.$$

Note that if subject i has entered the intermediate state, the conditional survival function in the intermediate state between times s and t is given by:

$$\frac{e^{-A_{12}(t|Z_{12i})}}{e^{-A_{12}(s|Z_{12i})}}.$$

We allow that the event times are left truncated, i.e., that subjects enter the study at the delayed entry time $E > 0$. The left truncation condition $X(E_i) = 0$ implies that subject i has survived in state 0 until time E_i. This is taken into account by dividing the above likelihood contributions by the term $S(E_i|Z_{01i}, Z_{02i})$.

In addition to the covariate vectors $Z_{01i}, Z_{02i}, Z_{12i}$, one for each transition, we observe the vector $(E_i, L_i, R_i, \delta_{1i}, \tilde{T}_i, \delta_{2i})$ where $\tilde{T}_i = \min(T_i, C_i)$ is the minimum between the transition time into the absorbing state T_i and the right censoring time C_i and $\delta_{2i} = I\{T_i \leq C_i\}$. Also, $\delta_{1i} = 1$ if we know for sure that subject i was diseased between E_i and $\tilde{T}_i$ and $\delta_{1i} = 0$ otherwise. The visit times L_i and R_i are defined by $E_i \leq L_i \leq R_i \leq \tilde{T}_i$ if $\delta_{1i} = 1$ and by $E_i \leq L_i \leq \tilde{T}_i, R_i = \infty$ if $\delta_{1i} = 0$. When the transition time into the intermediate state is observed exactly, we have $\delta_{1i} = 1$ and $L_i = R_i$. In the latter case we also denote I_i for the transition time into the intermediate state.

We now detail the likelihood contributions according to the different observational patterns shown in Figure 2:

$$\begin{aligned}
\text{case 1:}\quad \mathcal{L}_i &= S(C_i|Z_{01i}, Z_{02i}) + \int_{L_i}^{C_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(C_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}du \\
\text{case 2:}\quad \mathcal{L}_i &= \int_{L_i}^{R_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(C_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}du \\
\text{case 3:}\quad \mathcal{L}_i &= S(I_i|Z_{01i}, Z_{02i})\alpha_{01}(I_i|Z_{01i})\frac{e^{-A_{12}(C_i|Z_{12i})}}{e^{-A_{12}(I_i|Z_{12i})}} \\
\text{case 4:}\quad \mathcal{L}_i &= \int_{L_i}^{R_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}\alpha_{12}(T_i|Z_{12i})du \\
\text{case 5:}\quad \mathcal{L}_i &= S(I_i|Z_{01i}, Z_{02i})\alpha_{01}(I_i|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(I_i|Z_{12i})}}\alpha_{12}(T_i|Z_{12i}) \\
\text{case 6:}\quad \mathcal{L}_i &= S(T_i|Z_{01i}, Z_{02i})\alpha_{02}(T_i|Z_{02i}) \\
&\quad + \int_{L_i}^{T_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}\alpha_{12}(T_i|Z_{12i})du
\end{aligned} \tag{2}$$

3 Estimation

The idm function computes estimates for the three baseline transition intensities and for the regression parameters using the Levenberg-Marquardt's algorithm [9, 11] to maximize the (penalized) likelihood. The algorithm is a combination of a Newton-Raphson algorithm and a gradient descent algorithm (also known as the steepest descent algorithm). It has the avantage of being more robust than the Newton-Raphson algorithm while preserving its fast convergence property.

3.1 Parametric estimation

In the default estimation method of function idm, a Weibull parametrization for the baseline transition intensities is assumed:

$$\alpha_{0,hl}(t) = a_{hl}\ b_{hl}^{a_{hl}}\ t^{a_{hl}-1}; \quad hl \in \{01, 02, 12\}.$$

where a_{hl} and b_{hl} are shape and scale parameters. The Weibull parameters estimates $\hat{a}_{hl}$ and $\hat{b}_{hl}$ and the vectors of regression parameters estimates $\hat{\beta}_{hl}$ are obtained simultaneously by maximizing the likelihood which is the product over the subjects' contributions according to equation 2:

$$\mathcal{L}(\beta_{01}, \beta_{02}, \beta_{12}, a_{01}, a_{02}, a_{12}, b_{01}, b_{02}, b_{12}) = \prod_{i=1}^{n} \mathcal{L}_i(\beta_{01}, \beta_{02}, \beta_{12}, a_{01}, a_{02}, a_{12}, b_{01}, b_{02}, b_{12}).$$

Confidence intervals for the regression parameters are obtained using standard errors estimated by inverting the Hessian matrix of the log-likelihood, that is the matrix of the second partial derivatives of $\log \mathcal{L}$ given in the previous display. Pointwise confidence intervals for the baseline transition intensities are obtained using a simulation-based approach explained below (section 4.1).

3.2 Semi-parametric estimation

In situations where it is suspected that the Weibull distribution does not fit the data very well one can think of extending the model and to leave the baseline intensity functions completely unspecified, as in the Cox regression model. Unfortunately, in interval censored data there is no direct analogue to the partial likelihood and the Breslow estimator of the Cox model in right censored data. The function idm implements a semi-parametric model where the three baseline transition intensities are approximated by linear combinations of M-splines. In this section we explain the basic steps of the approach.

3.2.1 The penalized likelihood

To control the smoothness of the estimated intensity functions, we penalize the log-likelihood by a term which specificies the curvature of the intensity functions.

It is given by the square of the second derivates. The penalized log-likelihood (pl) is defined as:

$$pl = l - \kappa_{01} \int \alpha_{01}^{''\,2}(u|Z_{01})du - \kappa_{02} \int \alpha_{02}^{''\,2}(u|Z_{02})du - \kappa_{12} \int \alpha_{12}^{''\,2}(u|Z_{12})du \quad (3)$$

where l is the log-likelihood and κ_{01}, κ_{02} and κ_{12} are three positive parameters which control the trade-off between the data fit and the smoothness of the functions. It is proposed that the penalization parameters are chosen by maximizing a cross-validated likelihood score. Here, leave-one-out is appealing as the result does not depend on the random seed as it would, e.g., for 10-fold cross-validation. However, since leave-one-out requires as many maximizations of the likelihood as there are subjects in the data set, this can be computationally very expensive. To avoid extremely long run times we have implemented the following algorithm:

Step 1. We ignore the covariates and use a grid search method to find the values for $(\kappa_{01}, \kappa_{02}, \kappa_{12})$ based on an approximiation of the leave-one-out log-likelihood score. The approximation is equivalent to one step of the Newton-Raphson algorithm and reduces the number of calculations considerably. This approach was proposed by **(author?)** [12] for survival models and studied by **(author?)** [6] in an illness-death model with interval censored data.

Step 2. We use the results of Step 1, i.e. the optimized value of $(\kappa_{01}, \kappa_{02}, \kappa_{12})$ to maximize the penalized likelihood (3) with covariates. The parameters being maximized are the regression coefficients and the coefficients of the linear combination of the M-splines defined below.

3.2.2 M-splines

A family of M-spline functions of order k, $M_1, \ldots, M_n$ is defined by a set of m knots where $n = m + k - 2$ [13]. We consider only cubic M-splines of order $k = 4$. Denote by $t_{01} = (t_{01,1}, \ldots, t_{01,m_{01}})$ a sequence of m_{01} knots used for approximating α_{01} and by $t_{02} = (t_{02,1}, \ldots, t_{02,m_{02}})$ and $t_{12} = (t_{12,1}, \ldots, t_{12,m_{12}})$ similar sequences of knots for approximating α_{02} and α_{12} respectively. We denote by $M_{hl}^T = M_{hl,1}, \ldots, M_{hl,n_{hl}}$ the families of n_{hl} cubic M-splines, with $n_{hl} = m_{hl} + 2$ and for $hl \in \{01, 02, 12\}$. The baseline transition intensity $\alpha_{0,hl}$ is approximated using the following linear combination:

$$\tilde{\alpha}_{0,hl}(t) = \sum_{i=1}^{n_{hl}} (a_{hl,i})^2 M_{hl,i}(t)$$

where $a_{hl,i}$ are unkown parameters. The n_{hl} M-splines are integrated in order to produce a family of monotone splines, these are called I-splines. Thus, with each M-spline $M_{hl,i}$ we associate an I-spline $I_{hl,i}$:

$$I_{hl,i}(t) = \int_{t_{hl,1}}^{t} M_{hl,i}(u)du.$$

For given values of the parameters $a_{hl,i}$, we can approximate the cumulative baseline transition intensities A_{hl} by a linear combination of I-splines:

$$\tilde{A}_{0,hl}(t) = \sum_{i=1}^{n_{hl}} (a_{hl,i})^2 I_{hl,i}(t).$$

Because M-splines are non-negative, the positivity constraint on $(a_{hl,i})^2$ ensures that $\tilde{A}_{0,hl}$ is monotone increasing.

Confidence intervals of the regression parameters are obtained using estimated standard errors which are obtained by inverting the Hessian matrix of the penalized log-likelihood.

Confidence intervals for the transition intensities $\alpha_{hl}(t)$ are obtained using the Bayesian approach proposed in **(author?)** [12] for survival analysis where the standard errors are estimated by $M_{hl}(t)^T H^{-1} M_{hl}(t)$ where H denotes the Hessian matrix of the penalized log-likelihood.

4 Predictions

Often in illness-death models the functions of interest are the transition intensities. However, other quantities (transition probabilities, cumulative probabilities and life expectancies) which can be expressed in terms of the transition intensities [14] may provide additional information and have a more natural interpretation.

For example, given a set of covariates $Z_{01,i}, Z_{02,i}, Z_{12,i}$ for a subject i who is diseased at time s, one could be interested in probability to be still alive at some time $t > s$, or in life expectancy; given a set of covariates $Z_{01,j}, Z_{02,j}, Z_{12,j}$ for a subject j who is diseased-free at time s, one could be interested in lifetime risk of disease or in healthy life expectancy (expected remaining sojourn time in the healthy state). Since these quantities can be written in terms of the transition intensities, **SmoothHazard** provides estimates of them using estimates of the transition intensities. Confidence intervals are calculated using the simulation-based method immediately following.

4.1 Confidence regions

A simulation based approach [10] is used to calculate confidence intervals for the transition intensities $\alpha_{hl}(t)$ in the parametric approach and for the other quantities of interest in both parametric and semi-parametric approaches. To briefly outline how it works, we generically denote by θ the vector of all the parameters that characterize the likelihood and by $\hat{\theta}$ the maximum (penalized) likelihood estimator. θ contains the Weibull parameters in the parametric model, the spline parameters in the semi-parametric model and the regression parameters in both models.

We assume the asymptotic normality for the estimator $\hat{\theta}$. and denote by $\hat{V}_{\hat{\theta}}$ the estimated covariance matrix of $\hat{\theta}$. We consider a multivariate normal distribution with the parameters estimates as expectation and $\hat{V}_{\hat{\theta}}$ as covariance matrix. We generate n vectors ($n = 2000$ in practice) from this distribution: $\theta^{(1)}, \dots, \theta^{(n)}$. Based

Case	Description	δ_1	δ_2	L	R	T	Remark
1	No illness observed, right-censored death time	0	0	L_i	L_i	C_i	$L_i \leq C_i$
2	Interval-censored ill time, right-censored death time	1	0	L_i	R_i	C_i	$L_i < R_i \leq C_i$
3	Exact ill time, right-censored death time	1	0	L_i	L_i	C_i	$L_i \leq C_i$
4	Interval-censored ill time, death time observed	1	1	L_i	R_i	T_i	$L_i < R_i \leq T_i$
5	Exact ill time, death time observed	1	1	L_i	L_i	T_i	$L_i \leq T_i$
6	No illness observed, death time observed	0	1	L_i	L_i	T_i	$L_i \leq T_i$

Table 1: Description of how the data set must be built to be understood by the `idm` function

on them, we can calculate n values for the transition intensities: $\alpha_{hl}^{(1)}(t)$,...,$\alpha_{hl}^{(n)}(t)$, and therefore n values for any quantity of interest written in terms of the transition intensities. The n values reflecting the sample variation [1], we order them and the 2.5^{th} and the 97.5^{th} empirical percentiles are then used as lower and upper confidence bounds for 95% confidence intervals. This procedure can be repeated for any t, so we can obtain pointwise confidence limits for $\alpha_{hl}(.)$.

5 Using SmoothHazard

5.1 How to prepare the data

Table 1 shows how the program interpretes the structure of the data set. In all cases, L_i may be equal to the entry time. Some more details are necessary to distinguish the case where the ill status is known at the last follow-up time for death from the case where this is not possible.

- In case 1, if $L_i < C_i$ then it is assumed that the subject may become ill between L_i and C_i. If $L_i = C_i$ it is assumed that the subject is disease-free at time C_i. In the latter case the integral of the likelihood equals zero.
- In case 6, if $L_i < T_i$ then it is assumed that the subject may become ill between L_i and T_i. If $L_i = T_i$ it is assumed that the subject is disease-free at time T_i. In the latter case the of the likelihood equals zero.

5.2 Paquid study

In order to illustrate the functionality of the package we provide a random subset containing data from 1000 subjects that were enrolled in the Paquid study [8], a

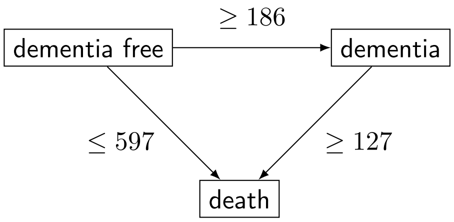


Figure 3: The exact number of transitions in the illness-death model with interval-censored time to disease is unknown.

large cohort study on mental and physical aging.

```r
library(SmoothHazard)
data(Paq1000)
```

The population consists of subjects aged 65 years and older living in Southwestern France. The event of interest is dementia and death without dementia is a competing risk. Furthermore, the time to dementia onset is interval censored between the diagnostic visit and the previous one and demented subjects are at risk of death. Thus, subjects who died without being diagnosed as demented at their last visit may have become demented between last visit and death.

In this subset 186 subjects are diagnosed as demented and 724 died from whom 597 without being diagnosed as demented before. Because of interval censoring more than 186 should have been demented, more than 127 should have been dead with dementia and less than 597 should have been dead without dementia (see Figure 5.2).

Age is chosen as the basic time scale and subjects are dementia-free (and alive) at entry into study. Consequently, we need to deal with left-truncated event times.

```r
head(round(Paq1000,1))
```

```
  dementia death    e    l    r    t certif gender
1        1     1 72.3 82.3 84.7 87.9      0      0
2        0     1 77.9 78.9 78.9 79.6      0      1
3        0     1 79.9 79.9 79.9 80.9      0      0
4        0     1 74.7 78.6 78.6 82.9      1      1
5        0     1 76.7 76.7 76.7 79.2      0      1
6        0     0 66.2 71.4 71.4 84.2      1      0
```

Each row in the data corresponds to one subject. The variables dementia and death are δ_1 and δ_2, the status variables for dementia and death. The variable `e` contains ages of subjects at entry into study. The variables `l` and `r` contain the left and right endpoints of the censoring intervals. For demented subjects, `r` is the age at the diagnostic visit and `l` is the age at the previous one. For non demented subjects, `l` and `r` are the age at the latest visit without dementia (`l`=`r`). The variable `t` is the age at death or at latest news on vital status. There are two binary covariates: certif for primary school diploma (762 with diploma and 238 without diploma) and gender (578 women and 422 men).

The function `idm` computes estimates for the three transition intensities $\alpha_{01}(.)$, $\alpha_{02}(.)$, $\alpha_{12}(.)$ which represents age-specific incidence rate of dementia, age-specific mortality rate of dementia-free subjects and age-specific mortality rate of demented subjects, respectively. Proportional transition intensities regression models allow for covariates on each transition. Covariates are specified independently for the regression models of the three transition intensities by the right hand side of the respective formula `formula01`, `formula02` and `formula12`.

Interval censoring and left truncation must be specified at the left side of the formula arguments using the `Hist` function. For left-truncated data, the entry argument of `Hist` must contain the vector of delayed entry times. For interval-censored data, the `time` argument of `Hist` must contain a list of the left and right endpoints of the intervals. The `data` argument contains the data frame in which to interpret the variables of `formula01`, `formula02` and `formula12`. The left side of `formula12` argument does not need to be filled because all the data informations are already contained in `formula01` and `formula02`. The left side of `formula12` argument is required only if we want the covariates impacting transition $1 \to 2$ different from those impacting transition $0 \to 2$.

5.3 Fitting the illness-death model based on interval-censored data

The main function `idm` computes estimates for the three baseline transition intensities and for the regression parameters of an illness-death model. The `intensities` argument by specifying the form of the transition intensities allows to select either the parametric or a semi-parametric estimation method :

- With the default value "Weib", a Weibull distribution is assumed for the baseline transition intensities and the parameters are estimated by maximizing the log-likelihood;
- With the "Splines" value, the baseline transition intensities are approximated by linear combinations of M-splines and the parameters are estimated by maximizing the penalized log-likelihood.

We stop the iterations of the maximization algorithm when the differences between two consecutive parameters values, log-likelihood values, and gradient values is small enough. The default convergence criteria are 10^{-5}, 10^{-5} and 10^{-3} and can be changed by means of the eps argument.

We now illustrate how to fit the illness-death model to the `Paq1000` data set, based on interval-censored dementia times and exact death times.

In the following call, a Weibull parametrization is used for the three baseline transition intensities and we include two covariates on the transition to dementia, one covariate on the transition from no dementia to death and no covariates on the transition from dementia to death. Note that in case of missing formula12 argument the covariates on the $1 \to 2$ transition are the same as the ones specified in the formula02 argument.

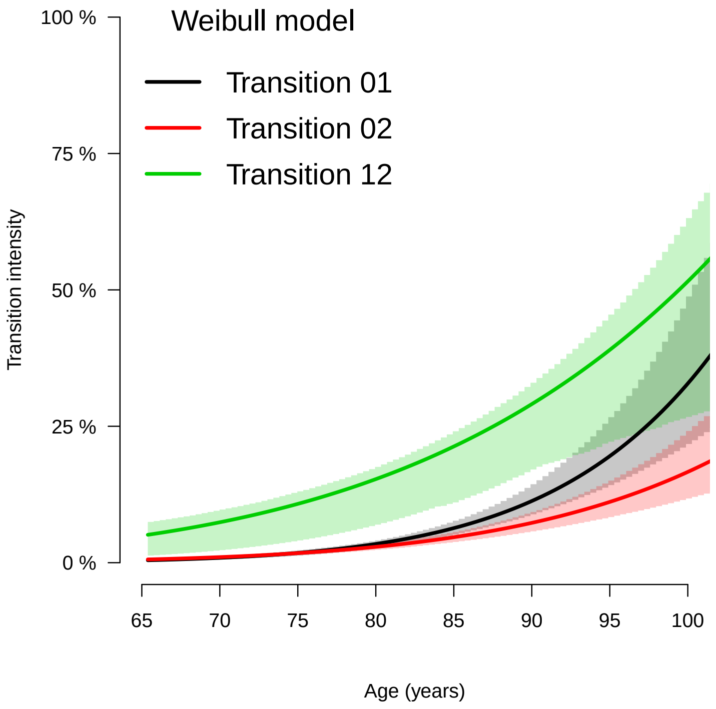
```
fit.weib <- idm(formula01=Hist(time=list(l,r),event=dementia,entry=e)~certif+gender,
        formula02=Hist(time=t,event=death,entry=e)~gender,
        formula12= ~ 1,
        data=Paq1000)
fit.weib
```

```
Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = e) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = e) ~ gender, formula12 = ~1, data = Paq1000)

Illness-death model: Results of Weibull regression for the intensity functions.

number of subjects:  1000
number of events '0-->1':  186
number of events '0-->2' or '0-->1-->2':  724
number of covariates:  2 1 0

                coef SE.coef     HR          CI      Wald p.value
certif_01_01 -0.4117  0.1827 0.6625 [0.46;0.95]  5.077106 0.02424
gender_01_01 -0.2621  0.1561 0.7694 [0.57;1.04]  2.818364 0.09319
gender_02_02  0.6712  0.1143 1.9565 [1.56;2.45] 34.449583 < 1e-04

               Without cov  With cov
Log likelihood   -3075.308 -3053.648

Parameters of the Weibull distribution: 'S(t) = exp(-(b*t)^a)'
      alpha01    alpha02    alpha12
a 11.12344625 8.82268159 6.44006486
b  0.01102198 0.01074539 0.01381268

----
Model converged.
number of iterations:  6
convergence criteria: parameters= 7.3e-10
                    : likelihood= 2.3e-08
                    : second derivatives= 2.8e-12
```

The hazard ratios HR (e^{coef}) have the usual interpretation, as in a parametric Cox regression model.

The three baseline transition intensity functions can be displayed as functions of time, functions of age in our illustrative example (Figure 3).

```
par(mgp=c(4,1,0),mar=c(5,5,5,5))
plot(fit.weib,conf.int=TRUE,lwd=3,citype="shadow",xlim=c(65,100),axis2.las=2,axis1.at=seq(65,100,5),xlab="Age (years)")
```



The other estimation option in the function idm permits to relax the strict parametric assumptions of the Weibull regression models. With the option intensities="Splines", linear combinations of M-splines are used to approximate the three baseline transition intensities. Although this option implies a considerable amount of extra computations (see Section 3.2), the call and the printed output are very similar to the Weibull model:

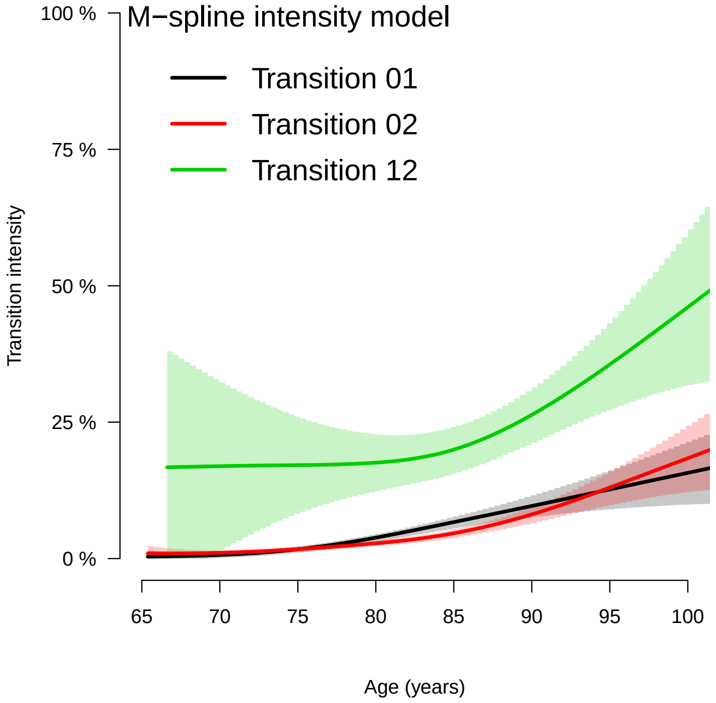
```
fit.splines <- idm(formula01=Hist(time=list(l,r),event=dementia,entry=e)~certif+gender,
        formula02=Hist(time=t,event=death,entry=e)~gender,
        formula12= ~ 1,
```

```r
         intensities="Splines",data=Paq1000)
fit.splines
```

```
Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = e) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = e) ~ gender, formula12 = ~1, data = Paq1000, intensities = "Splines")

Illness-death regression model using M-spline approximations
 of the baseline transition intensities.

number of subjects:  1000
number of events '0-->1':  186
number of events '0-->2' or '0-->1-->2':  724
number of subjects:  1000
number of covariates:  2 1 0

Smoothing parameters:
      transition01 transition02 transition12
knots        7e+00        7e+00            7
kappa        8e+05        2e+05        50000

                coef SE.coef     HR          CI      Wald p.value
certif_01_01 -0.3762  0.1853 0.6865 [0.48;0.99]  4.122728 0.04231
gender_01_01 -0.2297  0.1580 0.7948 [0.58;1.08]  2.113669 0.14599
gender_02_02  0.6529  0.1119 1.9211 [1.54;2.39] 34.039816 < 1e-04

                         Without cov  With cov
Penalized log likelihood   -3072.464 -3052.046

----
Model converged.
number of iterations:  8
convergence criteria: parameters= 4e-09
                    : likelihood= 9.5e-08
                    : second derivatives= 2.2e-11
```

Again, the estimated baseline transition intensities can conveniently be visualized in a joint graph (Figure 4).

```r
par(mgp=c(4,1,0),mar=c(5,5,5,5))
plot(fit.splines,conf.int=TRUE,lwd=3,citype="shadow",xlim=c(65,100),
    axis2.las=2,axis1.at=seq(65,100,5),xlab="Age (years)")
```



5.3.1 Semi-parametric estimation method: choice of smoothing parameters

Some optional arguments are specific to the semi-parametric approach (when using the option intensities="Splines"):

- n.knots contains a vector (by default c(7,7,7)) specifying the number of knots on the $0 \to 1$, $0 \to 2$ and $1 \to 2$ transitions, respectively;
- knots contains the choice of the knots placement (equidistant by default or quantile-based placement) or a list of sequences of knots for transitions $0 \to 1$, $0 \to 2$ and $1 \to 2$ respectively, to be specified by the user;
- CV (FALSE by default) is set to TRUE for using approximate leave-one-out cross-validation score to choose the smoothing parameters κ_{01}, κ_{02}, κ_{12};
- kappa contains the smoothing parameters if CV=FALSE (arbitrary choice of the smoothing parameters κ_{01}, κ_{02}, κ_{12}); the initial smoothing parameters for the grid search method which maximize the approximate leave-one-out cross-validation score if CV=TRUE.

By default the function idm selects equidistant sequences of 7 knots between the minimal and maximal event times (e, l and r for Paq1000). There must be a knot before or at the first time from which there are subjects at risk and after or at the last time of transition. The current implementation of our program requires a minimum of 4 knots for each transition intensitiy.

Consequently, the semi-parametric approach requires much more information than the parametric one to achieve convergence. The number of parameters to be estimated is larger, and enough observation times on each transition are required to fit the splines. In particular, in data sets where few $1 \to 2$ transitions times are observed, we does not recommended this approach. Increasing the number of knots does not deteriorate the estimates of the transition intensities: this is because the degree of smoothing in the penalized likelihood method is tuned by the smoothing parameters κ_{01}, κ_{12} and κ_{02}. On the other hand, once a sufficient number of knots is established, there is no advantage in adding more. Moreover, the more knots, the longer the running time. Some numerical problem can arise, particularly for a large number of knots. So it is recommended to start with a small number of knots (e.g. 5 or 7) and increase the number of knots until the graph of the transition intensities function remains unchanged (from our own experience rarely more than 12 knots).

The default values for the smoothing parameters κ_{01}, κ_{02}, κ_{01}, are suitable for the Paq1000 data set. However, these values can be expected to be very different depending on time scale, number of subjects and number of knots. The cross-validation option can be used to find appropriate smoothing parameters. However, the running time with cross-validation is very long and an empirical technique can be preferred. It consists in repeating the idm running trying different smoothing parameters. After each estimation, the transition intensities are plotted. If the curves seem too smooth, it may be useful to reduce the smoothing parameter. Similarly, if the curves are to wiggly, the smoothing parameter may be increased.

5.4 Making predictions

A object as returned by the idm function can be used as argument of the predict function in order to obtain transition probabilities, cumulative probabilities of event and life expectancies with confidence intervals. For example, the following call give predictions regarding a 70 years-old male subject who have primary school diploma, over a 10 years horizon:

```
pred <- predict(fit.weib,s=70,t=80,Z01=c(1,1),Z02=1)
x<-round(do.call("rbind",pred),2)
colnames(x) <- c("Probability","Lower","Upper")
x
```

```
      Probability Lower Upper
p00          0.64  0.59  0.68
p01          0.05  0.03  0.07
p11          0.33  0.27  0.67
p12          0.67  0.33  0.73
```

```
p02_0          0.29  0.24  0.33
p02_1          0.03  0.01  0.05
p02            0.32  0.27  0.36
F01            0.08  0.05  0.12
F0.            0.36  0.32  0.41
```

The covariates values must be specified in the Z01, Z02 and Z12 arguments in the same order as they were entered in the preceding idm call.

The ouput attributes are:

- for a dementia-free 70 years-old subject:
  - the probability of being still alive and dementia-free 10 years later $p_{00}(70, 80)$,
  - the probability of being still alive but demented 10 years later $p_{01}(70, 80)$,
  - the probability of dying in the next 10 years $p_{02}(70, 80)$ having been demented before $(p^1_{02}(70, 80))$ or not $(p^0_{02}(70, 80))$,
  - the absolute risk of dementia in the 10 years (10 years later, the subject may be dead or not) $F_{01}(s, t)$,
  - the absolute risk of exit from the no dementia state in the 10 years $F_{0\bullet}(s, t)$ (due to either dementia or death);
- for a demented 70 years-old subject: the probability of dying in the next 10 years $p_{12}(s, t)$ or not $p_{11}(s, t)$.

The following calls give life expectancies regarding a 80 years-old female subject who have primary school diploma based on the transition intensities estimates from respectively the parametric approach and the semi-parametric approach:

```r
LE.weib <- lifexpect(fit.weib,s=80,Z01=c(1,0),Z02=0)
x<-round(do.call("rbind",LE.weib),2)
colnames(x) <- c("LE","Lower","Upper")
x
```

```
                          LE Lower Upper
life.in.0.expectancy    8.87  7.89  9.78
life.expectancy.nondis 10.45  9.79 11.61
life.expectancy.dis     4.89  4.40  7.87
```

```r
LE.splines <- lifexpect(fit.splines,s=80,Z01=c(1,0),Z02=0,CI=FALSE)
x<-round(do.call("rbind",LE.splines),2)
colnames(x) <- c("LE")
x
```

```
                          LE
life.in.0.expectancy    8.82
life.expectancy.nondis 10.42
life.expectancy.dis     4.91
```

The confidence intervals calculation may take time, especially using the splines estimates of the transition intensities. To suppress this calculation, the CI argument must be set to FALSE (see above). To reduce the computation time of the confidence intervals, the number of simulations can also be modified using the nsim argument (by default 2000 for the predict function and 1000 for the lifexpect function).

The output attributes of the lifexpect function are:

- for a dementia-free 80 years-old subject:
  - the life expectancy in state 0 (healthy life expectancy),
  - the life expectancy;
- for a demented 80 years-old subject: the life expectancy.

5.4.1 Warnings regarding predictions

Predictions using the splines estimates of the transition intensities are not possible if involving times prior to the first knot or times beyond the last knot. Moreover, the life expectancies are calculated using integration until infinity using the Weibull estimates and until the last knot using the splines estimates. Consequently, to calculate life expectancies using the splines estimates, we implicitly assume that the last knot time is the maximal time of death. The above life expectancies calculating from the Weibull estimates or the splines estimates of the transition intensities are very close because the follow-up period of the Paq1000 data set is long. However, in other data sets this assumption may not hold anymore. For data sets with short follow-up period, it is possible to calculate quantities involving any time, even infinity like life expectancies. However, beyond the follow-up time, they are not based anymore on estimations of the transition itensity functions but rather on extrapolations on them. Consequently, we do not recommend to do predictions involving times beyond the follow-up period. Finally, to avoid numerical problem in the predictions calculations, the first and last knots for all transitions must be the same or very close.

References

[1] O O Aalen, V T Farewell, D De Angelis, N E Day, and O N Gill. A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to aids prediction in england and wales. *Statistics in Medicine*, 16(19):2191–2210, 1997.

[2] J. Beyersmann, A. Allignol, and M. Schumacher. *Competing Risks and Multistate Models with R*. Use R! Springer, 2011.

[3] D R Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

[4] Liesbeth C. de Wreede, Marta Fiocco, and Hein Putter. mstate: An r package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, 38(7):1–30, 2011.

[5] Christopher Jackson. Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, 38(8):1–28, 2011.

[6] P Joly, D Commenges, C Helmer, and L Letenneur. A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*, 3(3):433–443, 2002.

[7] K Leffondré, C Touraine, C Helmer, and P Joly. Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model? *International journal of epidemiology*, 2013.

[8] L Letenneur, V Gilleron, D Commenges, C Helmer, JM Orgogozo, and JF Dartigues. Are sex and educational level independent predictors of dementia and alzheimer's disease? incidence data from the paquid project. *Journal of Neurology, Neurosurgery & Psychiatry*, 66(2):177–183, 1999.

[9] Kenneth Levenberg. A method for the solution of certain problems in least squares. *Quarterly of applied mathematics*, 2:164–168, 1944.

[10] M Mandel. Simulation based confidence intervals for functions with complicated derivatives. *The American Statistician*, 67, 2013.

[11] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, 11(3):431–441, 1963.

[12] Finbarr O'Sullivan. Fast computation of fully automated log-density and log-hazard estimators. *Journal on Scientific and Statistical Computing*, 9(2):363–379, 1988.

[13] J O Ramsay. Monotone regression splines in action. *Statistical Science*, 3(4):425–441, 1988.

[14] Célia Touraine, Catherine Helmer, and Pierre Joly. Predictions in an illness-death model. *Statistical methods in medical research*, 2013.