Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Fitting regression models to interval-censored observations of illness-death models

Célia Touraine
University of Bordeaux

Thomas A. Gerds
University of Copenhagen

Pierre Joly
University of Bordeaux

Abstract

The irreversible illness-death model allows subjects to move from an initial state ("health state") to a terminal state ("death state") either directly or through an intermediate state ("disease state"). Disease onset times may not be known exactly, for example if the disease status of a patient can only be assessed at follow-up visits. In this situation the disease onset times are usually interval-censored. This article presents the **SmoothHazard** package for R. It implements algorithms for simultaneously fitting regression models to the three transition intensities of an illness-death model where the times to the intermediate state may be interval-censored data. The three baseline hazard functions are either parametrized according to Weibull distributions or approximated by M-splines. For a given set of covariates, the transition intensities estimates can be combined into predictions of transition probabilities, cumulative event probabilities and life expectancies.

Keywords: illness-death model, interval-censored data, left-truncated data, survival model, proportional regression models, smooth transition intensities, Weibull, penalized likelihood, M-splines.

1. Introduction

The irreversible illness-death model is a special multi-state model which has many applications for example in medical research. The model allows subjects to make transitions from an initial state (health) to a terminal state (death) either directly or via an intermediate state (disease), see Figure 1.

```
library(prodlim)
plotIllnessDeathModel(stateLabels=c("0: Healthy","1: Diseased","2: Dead"),arrowLabelSymbol="
    alpha")
```

If the exact transition times are observed, standard procedures like those implemented in the

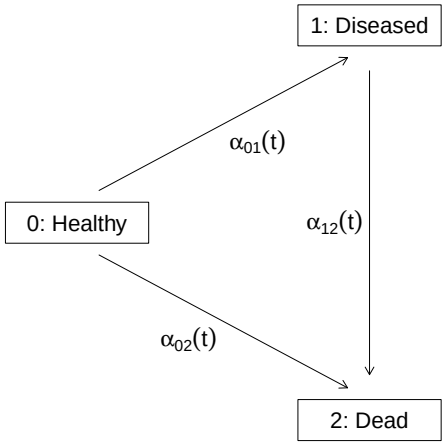

Figure 1: The irreversible illness-death model

mstate package can be used to estimate regression models for the three transitions (de Wreede *et al.* 2011). In particular, the regression coefficients can be estimated using Cox partial likelihood (Cox 1975) without the need to specify the baseline intensities. However, this possibility is lost when the transition times from the initial state to the intermediate state are interval censored. For example, it may not be possible to determine the exact onset time of disease for a subject diagnosed at time R. Instead it is only known that the subject was last seen disease-free at time L. Thus, the onset time is interval censored between L and R. Furthermore, for a subject who died without being diagnosed as diseased before, it may not be possible to determine if the subject developed disease between the last time he was seen and the time of death. Even if time-to-disease only is of interest, the risk of death should not be ignored using classical survival models, especially if risk of death for diseased subjects is higher than risk of death for disease-free subjects. A usual circumvention technique to handle subjects who died without being diagnosed as diseased before, consists in right-censoring dead subjects without disease diagnostic at the last time they were seen without disease. But this approach can lead to an underestimation of the transition intensity of disease (corresponding to the hazard function in survival models settings) (Joly *et al.* 2002) and biases in the regression coefficients estimates (LeffondrÃl' *et al.* 2013).

The **msm** package (Jackson 2011) allows to fit multi-state models to panel data (where states of each subjects are only known at a finite series of times) and could be used to fit illness-death models to interval-censored data. However, the illness-death model for interval-censored data is very simple compared to more general multi-state models for panel data. Consequently, the direct expression of the likelihood can be derived making possible to estimate more flexible transition intensities with less assumptions.

The algorithms of the **SmoothHazard** package implement methods for estimating regression models under interval censoring if the transition times into the absorbing state (e.g. death of the subject) are either known exactly or right censored (Joly *et al.* 2002). Implemented are a parametric and a semi-parametric estimation approach. For the parametric approach, the Weibull distribution is used for the baseline transition intensities and the parameters are estimated by maximising the likelihood. For the semi-parametric approach, M-splines are used to approximate the baseline transition intensities and the parameters are estimated

maximising a penalized likelihood. The methods allow delayed entry of the subjects, i.e. that the event times are left truncated.

Section 2 presents the model and the likelihood. Section 3 presents the estimation methods. Section 4 provides some examples illustrating **SmoothHazard** functions.

2. Data, model and questions

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 (Letenneur *et al.* 1999), a large cohort study on mental and physical aging.

```
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 2). There are two covariates in this subset: gender (`578` women and `422` men) and primary school diploma (`762` with diploma and `238` without diploma). 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.

```
head(Paq1000)
```

```
  dementia death      t0        l        r        t certif gender
1        1     1 72.3333 82.34014 84.73303 87.93155      0      0
2        0     1 77.9167 78.93240 78.93240 79.60048      0      1
3        0     1 79.9167 79.91670 79.91670 80.92423      0      0
4        0     1 74.6667 78.64750 78.64750 82.93501      1      1
5        0     1 76.6667 76.66670 76.66670 79.16636      0      1
6        0     0 66.2500 71.38070 71.38070 84.16975      1      0
```

Each row in the data corresponds to one subject. The variables `dementia` and `death` are the status variables (1 if an event occured, 0 otherwise) for dementia and death, respectively. The variable `t0` 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. `certif` and `gender` are binary covariates.

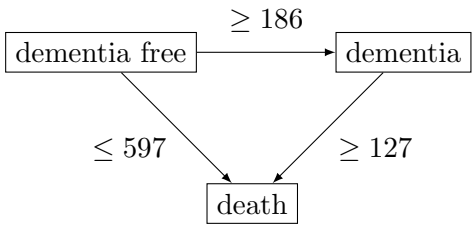

Figure 2: The exact number of transitions in the illness-death model with interval-censored time to disease is unknown.

The function `idm` computes estimates for the three transition intensities $\alpha_{01}(.)$, $\alpha_{02}(.)$, $\alpha_{12}(.)$ which are age-specific incidence rate of dementia, age-specific mortality rate of dementia-free subjects and age-specific mortality rate of demented sujects, 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.

A call of the `idm` function looks as follows:

```
fit <- idm(formula01=Hist(time=list(l,r),event=dementia,entry=t0)~certif,
           formula02=Hist(time=t,event=death,entry=t0)~certif+gender,
           formula12= ~ 1,
           data=Paq1000)
```

where the `data` argument contains the data frame in which to interpret the variables of `formula01`, `formula02` and `formula12`.

Note that the left side of `formula12` does not need to be filled because all the data informations are already contained in `formula01` and `formula02`. In fact, the `formula12` argument is required only if we want the covariates impacting transition 12 different from those impacting transition 02.

Questions ? TODO

3. Fitting the illness-death model based on interval-censored data

We consider an illness-death process $X = (X(t), t \geq 0)$. $X(t)$ has values in $\{0, 1, 2\}$. Subjects are initially dementia-free (state 0) and may become demented (transition $0 \to 1$) and die (transition $1 \to 2$) or, die directly (transition $0 \to 2$.) 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 $X(s)$. X is fully characterized by the transition probabilities :

$$p_{hl}(s,t) = \mathbb{P}(X(t) = l | X(s) = h)$$

or the transition intensities which are instantaneous transition probabilities :

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

The transition intensities in multi-state models are similar to hazard functions in survival models.

We introduce covariates through proportional transition intensity models which are a natural extension of the Cox proportional hazard model :

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

where $\alpha_{0,hl}$ are baseline transition intensities, Z_{hli} are covariates vectors for subject i and β_{hl} are vectors of regression parameters.

In the situation where time to disease and time to death are not interval censored the regression coefficients could be estimated by the partial likelihood method without the need to specify or estimate the baseline hazard functions $\alpha_{0,hl}(t)$. For interval-censored transition times to state 1 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, let us denote the conditional event-free survival function by

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

where $A_{hl}(.|Z_{hli})$ are the conditional cumulative intensity functions:

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

Age is chosen as the basic time scale and the model assumes that subjects are dementia-free at entry into the cohort. Let us denote T_{0i} the age of subject i at entry into the cohort. The left truncation condition $X(T_{0i}) = 0$ is taken into account into the likelihood by the term $\frac{1}{S(T_{0i}|Z_{01i}, Z_{02i})}$. Note that without left truncation data, $T_{0i} = 0$ and this term would disappear.

We set $\delta_{1i} = 1$ ($\delta_{1i} = 0$) if subject i has (has not) been observed diseased, and $\delta_{2i} = 1$ ($\delta_{2i} = 0$) if subject i is (is not) dead.

If $\delta_{2i} = 0$, T_i is time to death; if $\delta_{2i} = 0$, death event is right-censored at T_i. We denote by L_i and R_i the interval censoring times. If subject i has been observed diseased at time R_i and has last been seen non diseased at time L_i ($L_i < R_i$), time to disease is interval-censored between L_i and R_i. The likelihood contribution for subject i is:

$$\mathcal{L}_i = \frac{1}{S(T_{0i}|Z_{01i}, Z_{02i})} \int_{L_i}^{R_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i}) \frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}} \big(\alpha_{12}(T_i|Z_{12i})\big)^{\delta_{2i}} du \tag{2}$$

If subject i has never been seen diseased, time to disease is right-censored and the interval censoring times are set to the right censoring time ($L_i = R_i$). The likelihood contribution for subject i is:

$$\mathcal{L}_i = \frac{1}{S(T_{0i}|Z_{01i}, Z_{02i})}\left(S(T_i|Z_{01i}, Z_{02i})\big(\alpha_{02}(T_i|Z_{02i})\big)^{\delta_{2i}} + \int_{L_i}^{T_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}\big(\alpha_{12}(T_i|Z_{12i})\big)^{\delta_{2i}} du \right) \quad (3)$$

If time to disease and time to death are both right-censored at the same time, we have $L_i = R_i = T_i$ and the integral value in (3) is zero.

The `idm` function computes estimates for the three baseline transition intensities and for the regression parameters using likelihood-based estimation methods. In the parametric method and in the semi-parametric method, respectively the likelihood and the penalized likelihood are maximized using the Levenberg-Marquardt's algorithm (Levenberg 1944; Marquardt 1963) which is a combination of a Newton-Raphson algorithm and a gradient descent algorithm (also known as the steepest descent algorithm). This algorithm has the avantage of being more robust than the Newton-Raphson algorithm while preserving its fast convergence property. We stop the iterations 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. The variances of parameter estimates are estimated using the inverse of the matrix of the second derivatives at convergence.

3.1. The Weibull parametrization

The default estimation method in function `idm` computes the maximum likelihood estimate for the three transition intensities using a Weibull parametrization for the baseline transition intensities:

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

where a_{hl} and b_{hl} are shape and scale parameters. The Weibull parameters a_{hl} and b_{hl} and the vectors of regression parameter $\hat{\beta}_{hl}$ are obtained simultaneously by maximizing the log-likelihood.

```
fit.weib <- idm(formula02=Hist(time=t,event=death,entry=t0)~certif+gender,
                formula01=Hist(time=list(l,r),event=dementia,entry=t0)~certif+gender,
                data=Paq1000,intensities="Weib")
print(fit.weib)
```

```
Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = t0) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = t0) ~ certif + gender, data = Paq1000, intensities = "Weib")

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 2 2
number of deleted observations due to missing:  1

             coef SE.coef     HR          CI       Wald  p.value
certif_01 -0.5194  0.2015 0.5949 [0.40;0.88]  6.6405121 0.009969
gender_01 -0.1221  0.1599 0.8851 [0.65;1.21]  0.5832909 0.445025
certif_02  0.1268  0.1264 1.1352 [0.89;1.45]  1.0066554 0.315705
gender_02  0.5363  0.1200 1.7096 [1.35;2.16] 19.9877044 < 0.0001
certif_12 -0.2079  0.2322 0.8123 [0.52;1.28]  0.8016825 0.370591
gender_12  0.5792  0.1865 1.7846 [1.24;2.57]  9.6464868 0.001897

               Without cov  With cov
Log likelihood   -3075.308 -3048.791

Parameters of the Weibull distribution: 'S(t) = exp(-(b*t)^a)'
      alpha01    alpha02    alpha12
a 11.18802185 8.62750164 7.50200262
b  0.01099806 0.01078284 0.01294115

----
Model converged.
number of iterations:  8
convergence criteria: parameters= 0.00000012
                    : likelihood= 0.0000007
                    : second derivatives= 0.00000000047
```

The regression parameters HR have the usual interpretation, as in a fully parametric Cox regression model (CELIA, PIERRE: is this correct? There are some confusing other parametrizations of the Weibull model, eg. the function 'psm' of harrel's R-package rms)

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)")
```

Data were simulated to illustrate the effect of the length between visit times on the properties of the estimtated regression parameters in the Weibull model.

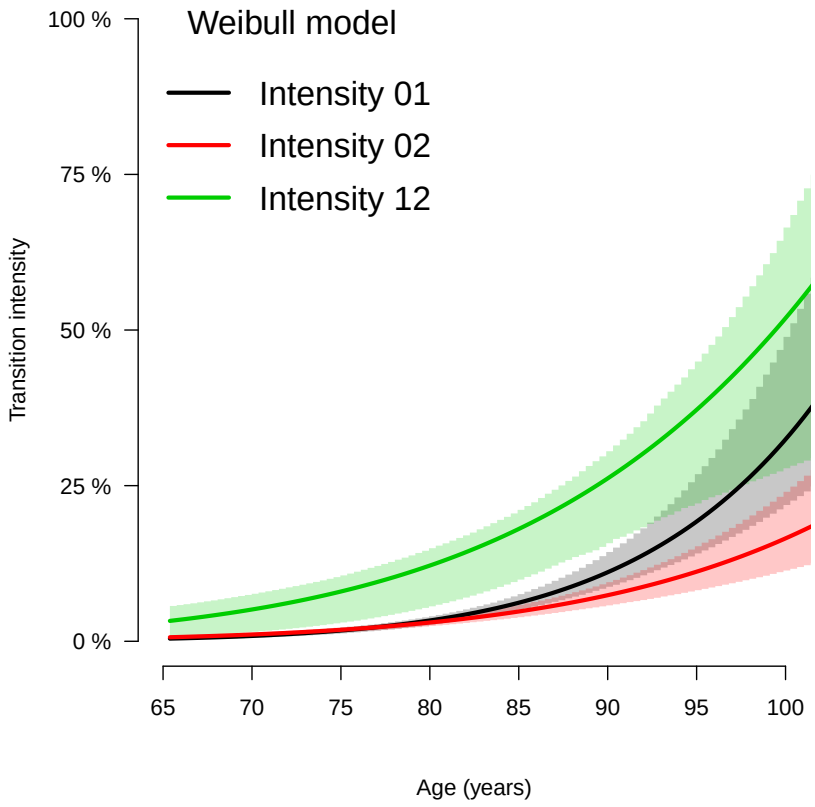

Figure 3: Estimated baseline intensities using Weibull regression for all transitions in the `Paq1000` data.

N	b01	b02	schedule	trans 0 -> 1	trans 0 -> 2	trans 1 -> 2
50	0.69	0.00	5	0.51 (0.4)	-0.26 (0.2)	-0.22 (0.3)
250	0.69	0.00	5	0.47 (0.3)	-0.25 (0.2)	-0.23 (0.2)
500	0.69	0.00	5	0.47 (0.3)	-0.24 (0.2)	-0.23 (0.2)
50	0.69	0.69	5	0.062 (0.09)	0.025 (0.08)	-0.249 (0.23)
250	0.69	0.69	5	0.051 (0.02)	0.019 (0.01)	-0.279 (0.10)
500	0.69	0.69	5	0.047 (0.008)	0.015 (0.006)	-0.285 (0.090)
50	0.69	0.00	35	0.54 (0.5)	-0.25 (0.2)	-0.28 (0.3)
250	0.69	0.00	35	0.48 (0.3)	-0.24 (0.2)	-0.29 (0.2)
500	0.69	0.00	35	0.48 (0.3)	-0.24 (0.2)	-0.29 (0.2)
50	0.69	0.69	35	0.068 (0.09)	0.034 (0.07)	-0.404 (0.31)
250	0.69	0.69	35	0.041 (0.01)	0.020 (0.01)	-0.423 (0.20)
500	0.69	0.69	35	0.039 (0.008)	0.015 (0.006)	-0.433 (0.198)

schedule	N	trans 0 -> 1	trans 0 -> 2	trans 1 -> 2	converged
0	100	-0.27 (0.2)	-0.23 (0.2)	-0.16 (0.3)	59.8 %
5	100	-0.207 (0.2)	-0.159 (0.2)	-0.019 (0.3)	69.5 %
20	100	0.0089 (0.04)	0.0694 (0.07)	0.1855 (0.18)	97.3 %
35	100	0.020 (0.04)	0.077 (0.06)	0.151 (0.16)	98.6 %
0	250	-0.26 (0.2)	-0.21 (0.2)	-0.19 (0.2)	62.1 %
5	250	-0.132 (0.1)	-0.063 (0.1)	0.022 (0.2)	81.3 %
20	250	0.011 (0.01)	0.081 (0.02)	0.142 (0.06)	100 %
35	250	0.012 (0.01)	0.082 (0.02)	0.101 (0.05)	100 %
0	500	-0.20 (0.1)	-0.15 (0.1)	-0.12 (0.2)	71 %
5	500	-0.0749 (0.06)	-0.0087 (0.06)	0.0788 (0.09)	89.3 %
20	500	0.009 (0.006)	0.072 (0.013)	0.120 (0.032)	100 %
35	500	0.0096 (0.006)	0.0723 (0.013)	0.0814 (0.025)	100 %

3.2. The penalized likelihood approach

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 below), the call and the printed output are very similar to the Weibull model:

```
fit.splines <- idm(formula02=Hist(time=t,event=death,entry=t0)~certif+gender,
                   formula01=Hist(time=list(l,r),event=dementia,entry=t0)~certif+gender,
                   data=Paq1000,intensities="Splines")
print(fit.splines)
```

```
Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = t0) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = t0) ~ certif + gender, data = Paq1000, intensities = "Splines")

Illness-death model using a penalized likelihood approach with splines
approximation 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 subjects:  1000
number of covariates:  2 2 2
number of deleted observations due to missing:  1

Smoothing parameters:
      transition01 transition02 transition12
knots            7            7            7
kappa      1000000       500000        20000
```

```
             coef SE.coef     HR          CI       Wald  p.value
certif_01 -0.4981  0.2075 0.6077 [0.40;0.91]  5.7622609 0.016374
gender_01 -0.0558  0.1655 0.9458 [0.68;1.31]  0.1135984 0.736084
certif_02  0.1290  0.1283 1.1376 [0.88;1.46]  1.0101740 0.314861
gender_02  0.5043  0.1215 1.6558 [1.30;2.10] 17.2238696 < 0.0001
certif_12 -0.2037  0.2388 0.8157 [0.51;1.30]  0.7273826 0.393733
gender_12  0.6449  0.1934 1.9058 [1.30;2.78] 11.1222626 0.000853

                         Without cov  With cov
Penalized log likelihood   -3073.099 -3046.848

----
Model converged.
number of iterations:  8
convergence criteria: parameters= 0.0000000076
                    : likelihood= 0.0000002
                    : second derivatives= 0.00000000005
```

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

```
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)")
```

3.3. Penalized likelihood

Let us denote the log-likelihood by l. 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 that is the square of the second derivates. The penalized log-likelihood (pl) is defined as:

$$pl = l - \kappa_{01} \int \alpha_{01}^{''\,2}(u|Z_{01})du - \kappa_{02} \int \alpha_{02}^{''\,2}(u|Z_{02})du - \kappa_{12} \int \alpha_{12}^{''\,2}(u|Z_{12})du \qquad (4)$$

where l is the log-likelihood and κ_{01}, κ_{02} and κ_{12} are three positive smoothing parameters which control the trade-off between the data fit and the smoothness of the functions. The smoothing parameters can be chosen either arbitrarily or by maximizing a cross-validation score. The leave-one-out cross-validation involves using a single observation as the validation data and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as validation data. To avoid maximizing the likelihood as many times as there are observations in the data set (and for each different values of κ_{01}, κ_{02}, κ_{12}), we use an approximate leave-one-out cross-validation score proposed by O'Sullivan (1988) for survival models and extended by Commenges *et al.* (2007) to multi-state models. It requires maximizing the likelihood one time only by tested values of the smoothing parameters. To find the values of κ_{01}, κ_{02}, κ_{12} which maximize this score we use a golden section search.

For given smoothing parameters, maximization of (4) defines the maximum penalized likelihood estimators of the baseline transition intensities $\hat{\alpha}_{0,01}$, $\hat{\alpha}_{0,02}$ and $\hat{\alpha}_{0,12}$ which are ap-

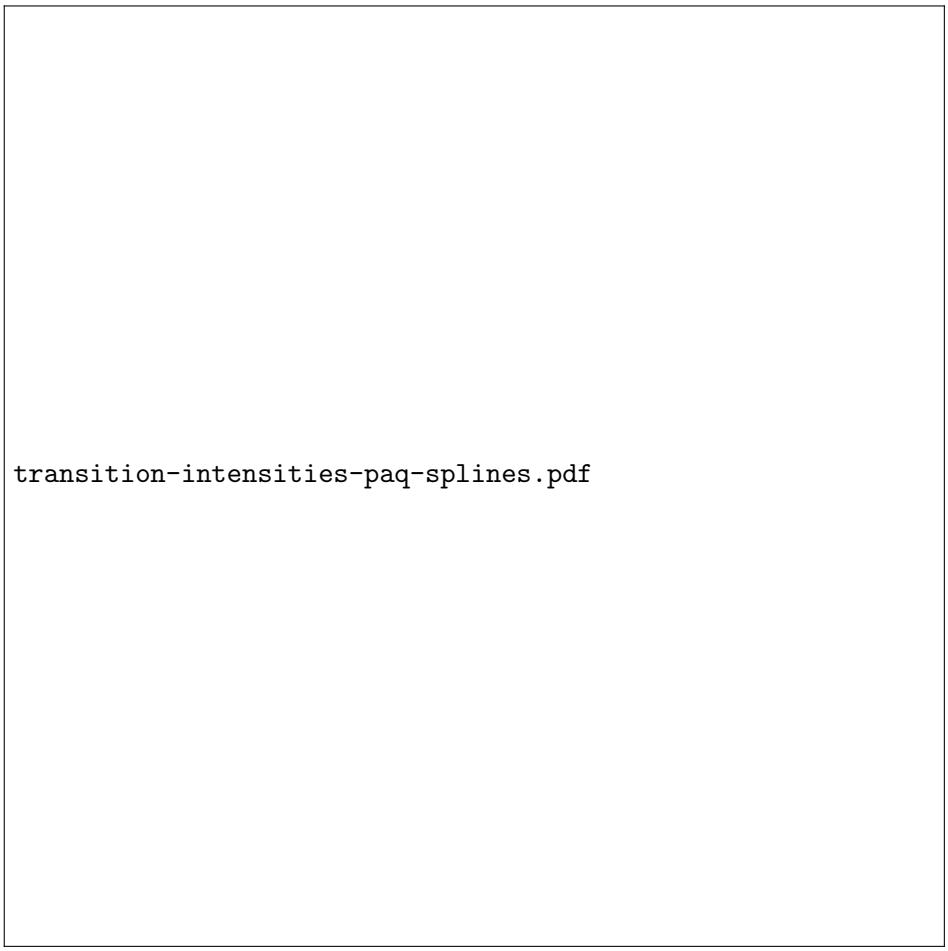

Figure 4: Estimated baseline intensities using M-splines for all transitions in the `Paq1000` data.

proximated using a basis of M-splines. The parameters being maximized are the regression coefficients and the coefficients of the linear combination of M-splines.

3.4. Splines approximation

A M-spline (see (Ramsay 1988)) is a non negative spline. A family of M-spline functions of order k, $M_1, \ldots, M_n$ is defined by a set of m knots $t = (t_1 \leq t_2 \leq \ldots \leq t_m)$ where $n = m + k - 2$. We use cubic M-splines, i.e. M-splines of order $k = 4$.

We denote by $t_{01} = (t_{01,1}, \ldots, t_{01,m_{01}})$ the sequence of m_{01} knots used to define the cubic M-splines approximation of $\hat{\alpha}_{0,01}$, 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 for $\hat{\alpha}_{0,02}$ and $\hat{\alpha}_{0,12}$, respectively. We denote by $M_{01,1}, \ldots, M_{01,n_{hl}}$ the matching family of n_{hl} cubic M-splines, with $n_{hl} = m_{hl} + 2$.

For $hl \in \{01, 02, 12\}$, the estimator $\hat{\alpha}_{hl}$ is approximated using the linear combination:

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

where $a_{hl,i}$ are the coefficients to estimate.

Non-negativity of $\tilde{\alpha}_{0,hl}$ is obtained by constraining the coefficients $a_{hl,i}$ to be positive. In practice, we estimate parameters $\theta_{hl,i}$ such that $a_{hl,i} = \theta_{hl,i}^2$.

The n_{hl} M-splines can be integrated to produce a family of monotone splines, called I-splines. With each M-spline $M_{hl,i}$ we associate an I-splines $I_{hl,i}$:

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

Given the coefficients $a_{hl,i}$, we can approximate the estimators of the cumulative baseline transition intensities $\hat{A}_{hl}$ by a linear combination of I-splines:

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

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

Choice of the knots

By default the function `idm` selects equidistant sequences of 7 knots. For the transition $h \to l$, the first knot is set to the minimal time from which there are subjects at risk of making the $h \to l$ transition and the last knot is set to the maximal time of the $h \to l$ transition times. For example, in the Paquid data set, the first knot for the transitions $0 \to 1$ and $0 \to 2$ is the minimal age of entry into the cohort and the first knot for the $1 \to 2$ transition is the minimal age of dementia of the subjects who have been observed demented. Since $0 \to 1$ times are interval-censored, the left bound of the interval is used. The last knot of the $0 \to 1$ transition is the maximal age of dementia (the right bound of the interval is used) and the last knots of the $0 \to 2$ and $1 \to 2$ transitions are the maximum death time. (PIERRE: isn'it the better way to choose the knots ? If yes, it must be improved in the package)

The placement of knots can be controlled by the argument `knots`. They are equidistant by default, but a quantile-based placement can also be chosen or the user can specify in a list sequences of its own knots in the order t^{01}, t^{02}, t^{12}. Generally the shape of a spline function is not very sensitive to knot placement. However, there must be several data points between each pair of different knots and 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.

Here are results with a different choice of knots:

```
#x <- sort(unique(unlist(Paq1000[,c("l","r","t0","t")])))
#hist(x)
fit.splines <- idm(formula02=Hist(time=t,event=death,entry=t0)~certif+gender,
                   formula01=Hist(time=list(l,r),event=dementia,entry=t0)~certif+gender,
                   data=Paq1000,intensities="Splines",
knots=list(c(65,75,80,82,84,86,88,90,105),c(65,75,80,82,84,86,88,90,105),c(65,80,90,105)))
print(fit.splines)
```

```
Call:
```

```
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = t0) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = t0) ~ certif + gender, data = Paq1000, knots = list(c(65,
    75, 80, 82, 84, 86, 88, 90, 105), c(65, 75, 80, 82, 84, 86,
    88, 90, 105), c(65, 80, 90, 105)), 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 2 2

Smoothing parameters:
      transition01 transition02 transition12
knots        9e+00        9e+00            4
kappa        8e+05        2e+05        50000

                coef SE.coef     HR          CI       Wald  p.value
certif_01_01 -0.4876  0.2075 0.6141 [0.41;0.92]  5.5246128 0.018751
gender_01_01 -0.0568  0.1656 0.9448 [0.68;1.31]  0.1177033 0.731539
certif_02_02  0.1190  0.1275 1.1264 [0.88;1.45]  0.8721192 0.350369
gender_02_02  0.5048  0.1202 1.6566 [1.31;2.10] 17.6292130  < 1e-04
certif_12_12 -0.1982  0.2384 0.8202 [0.51;1.31]  0.6913637 0.405701
gender_12_12  0.6135  0.1913 1.8469 [1.27;2.69] 10.2912825 0.001337

                        Without cov  With cov
Penalized log likelihood   -3072.556 -3047.102

----
Model converged.
number of iterations:  9
convergence criteria: parameters= 1.5e-07
                    : likelihood= 3.2e-06
                    : second derivatives= 6.6e-11
```

The number of knots can be controlled by the argument `n.knots` of the function `idm`. By increasing the number of data points between a pair of knots, i.e. by selecting fewer knots, one achieves a more smooth but less flexible approximation (CELIA, PIERRE: this last statement is just my intuition, is it correct? you wrote: "Increasing the number of data points between a pair of knots leads to a better defined curve." I did not understand what you meant.)

Increasing the number of knots does not deteriorate the MPLE: 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. That is why the maximum number of knots is limited to 25. 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 (rarely more than 12 knots).

Choice of smoothing parameters

The default values for the smoothing parameters are suitable for the `Paq1000` data set. However, these values can be expected to be very different depending on time scale and number of subjects. They can be changed into the `kappa` argument. The can also be chosen using by cross-validation using the argument `CV` (FALSE by default). In this case, the `kappa` argument contains the initial values for golden section search of the 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. This can be done with the `plot` function. If the curves seem too smooth it may be useful to reduce the associated smoothing parameter. Similarly, if the curves are to wiggly, the associated smoothing parameter may be increased.

4. Predicting parameters of life

Most often in illness-death models, the functions of interest are the transition intensities. However, other functions/quantities which can be expressed in terms of the transition intensities (Touraine *et al.* 2013) and may provide additional information and have a more natural interpretation.

The function `idm` returns an "idmWeib" or "idmSplines" class object depending on the parametrization of the transition intensities (Weibull or splines). These objects can be used in argument of the `predict.idmWeib` and `predict.idmSplines` functions in order to obtain transition probabilities between ages 70 and 80 (and cumulative probabilities). For example, for a female subject who is healthy at age 70 and has primary school certificate:

```
TP <- predict(fit.weib,s=70,t=80,Z01=c(1,0),Z02=c(1,0),Z12=c(1,0))
TP
```

```
$p00
[1] 0.8899268 0.8245325 0.8759390

$p01
[1] 0.03041876 0.02406653 0.05422249

$p11
[1] 0.2649379 0.2802728 0.7980570

$p12
[1] 0.7350621 0.2019430 0.7197272

$p02_0
```

```
[1] 0.06035533 0.07735229 0.12578106

$p02_1
[1] 0.01929910 0.00399737 0.02334086

$p02
[1] 0.07965443 0.09042787 0.13498797

$F01
[1] 0.04971786 0.03197171 0.07140308

$F0.
[1] 0.1100732 0.1240610 0.1754675
```

where `TP$p00`, `TP$p01`, `TP$p11`, `TP$p02` are the transition probabilities; `TP$p02_1` and `TP$p02_0` are the probabilities of transition from state 0 to state 2 coming through state 1 or not; `F01` is the probability for of becoming diseased between ages 70 and 80; `F0.` is the probability of exit from state 0 between ages 70 and 80.

If the `predict` function is used with an `idmSplines` object, the `s` input must be greater than the first knot and the `t` input must be lower than the last knot.

The "idmWeib" or "idmSplines" objects can also be used in argument of the `lifexpect` function to obtain life expectancies. For example, for a female subjects who has primary school certificate, the following code:

```
LE_fit.weib <- lifexpect(fit.weib,s=90,Z01=c(1,0),Z02=c(1,0),Z12=c(1,0),CI=FALSE)
LE_fit.weib
```

```
$life.in.0.expectancy
[1] 4.047053

$life.expectancy.nondis
[1] 5.416446

$life.expectancy.dis
[1] 4.057387
```

produces healthy life expectancy, life expectancy for a non diseased subject and life expectancy for a diseased subject. Again, if this function is used with an `idmSplines` object, the `s` input must be greater than the first knot. Moreover, life expectancies are calculated integrating up to infinity using an "idmWeib" object but up to the last knot using an "idmSplines" object. Consequently using an "idmSplines" object, it must be acceptable to assume that any subject should be in state 2 (dead) at the age corresponding to the last knot. Otherwise, the life expectancies would be underestimated.

References

Commenges D, Joly P, Gégout-Petit A, Liquet B (2007). "Choice between Semi-parametric Estimators of Markov and Non-Markov Multi-state Models from Coarsened Observations." *Scandinavian Journal of Statistics*, **34**(1), 33–52.

Cox DR (1975). "Partial Likelihood." *Biometrika*, **62**, 269–276.

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.

Jackson C (2011). "Multi-State Models for Panel Data: The msm Package for R." *Journal of Statistical Software*, **38**(8), 1–28.

Joly P, Commenges D, Helmer C, Letenneur L (2002). "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.

LeffondrÃl' K, Touraine C, Helmer C, Joly P (2013). "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*.

Letenneur L, Gilleron V, Commenges D, Helmer C, Orgogozo J, Dartigues J (1999). "Are sex and educational level independent predictors of dementia and AlzheimerâĂŹs disease? Incidence data from the PAQUID project." *Journal of Neurology, Neurosurgery & Psychiatry*, **66**(2), 177–183.

Levenberg K (1944). "A method for the solution of certain problems in least squares." *Quarterly of applied mathematics*, **2**, 164–168.

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

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

Ramsay JO (1988). "Monotone regression splines in action." *Statistical Science*, **3**(4), 425–441.

Touraine C, Helmer C, Joly P (2013). "Predictions in an illness-death model." *Statistical methods in medical research*.

Affiliation:

Célia Touraine
Univ. Bordeaux
ISPED
Centre INSERM U-897-Epidemiologie-Biostatistique
Bordeaux F-33000
France
E-mail: celia.touraine@isped.u-bordeaux2.fr
URL: http://www.isped.u-bordeaux2.fr/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd