Graphical Abstract

Penalised maximum likelihood estimation in multi-state models for interval-censored data

Robson J. M. Machado, Ardo van den Hout, Giampiero Marra

Highlights

Penalised maximum likelihood estimation in multi-state models for interval-censored data

Robson J. M. Machado, Ardo van den Hout, Giampiero Marra

- A new and efficient method to estimate multi-state models with splines using automatic estimation of penalty parameters.
- A simulation study and two data analyses illustrate the method for interval-censored multi-state data.

Penalised maximum likelihood estimation in multi-state models for interval-censored data

Robson J. M. Machado, Ardo van den Hout*, Giampiero Marra

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

Abstract

Continuous-time multi-state Markov models can be used to describe transitions over time across health states. Given longitudinal interval-censored data on transitions between states, statistical inference on changing health is possible by specifying models for transition hazards. Parametric time-dependent hazards can be restrictive, and nonparametric hazard specifications using splines are presented as an alternative. The smoothing of the splines is controlled by using penalised maximum likelihood estimation. With multiple time-dependent hazards in a multi-state model, there are multiple penalty parameters and selecting the optimal amount of smoothing is a challenge. A grid search to estimate the penalty parameters is computational intensive especially when combined with methods to deal with interval-censored transition times. A new and efficient method is proposed to estimate multi-state models with splines where the estimation of the penalty parameters is automatic. A simulation study is undertaken to validate the method and to illustrate the effect of interval censoring. The feasibility of the method is illustrated with two applications.

Keywords: Automatic smoothing, Markov model, Panel data, Penalised splines, Survival analysis

*Corresponding author. Tel: +44(0)20 3108 3243. Fax: +44 (0)20 3108 3105.
Email address: ardo.vandenhout@ucl.ac.uk (Ardo van den Hout)

1. Introduction

In biostatistics, disease progression can often be investigated using longitudinal data on change of health status. Multi-state models are commonly used to describe transitions across a set of discrete states. When time of transitions are observed intermittently, the data on the transition times are interval-censored. For continuous-time multi-state models, a time-homogeneous Markov process is usually assumed (Kalbfleisch and Lawless, 1985; Jackson, 2011). For a wide range of applications, however, the risks of moving across states depend on the current state and on time. In this case, a non-homogeneous Markov assumption is assumed to model the multi-state process.

Several time-dependent models can be fitted with parametric specifications for transition hazards (Cook and Lawless, 2018; Van den Hout, 2017). However, the functional form describing the transition hazards as a function of time is often unknown and parametric models can be too restrictive.

Alternatively, splines can be used to model the time dependency of transition hazards. Splines are piecewise polynomial functions, and a semiparametric hazard model is defined by a weighted sum of basis functions, where the weights in the sum are parameters that have to be estimated. It is possible to define a spline using many basis functions, and this will allow for a flexible model across the whole time range in the data. Penalised maximum likelihood estimation can be used to estimate parameters. This estimation includes a smoothing (or penalty) parameter that balances smoothness of the fitted hazard across the whole time range against fidelity to the data.

A penalised maximum likelihood estimation for a progressive three-state model is developed in Joly and Commenges (1999). Estimation is performed with an algorithm which uses analytical derivatives of the penalised log-likelihood. The smoothing parameters are selected using a grid search with cross-validation. In this case, models have to be fitted for every combination of smoothing parameters defined by the grid. Joly et al. (2002) use the same method for an illness-death model. This method can be computationally intensive for models with multiple smoothing parameters; that is, for models where multiple transition hazards each have their own smoothing parameter. In addition, the method requires explicit expressions for the transition probabilities. Calculating those formulae can be intractable for more complex models, such as models with more than four states and backwards transitions (Jackson, 2011). Titman (2011) uses a numerical approxima-

tion to calculate the transition probabilities at the level of the corresponding differential equations. The method allows for nonparametric hazard specifications with B-spline basis functions placed equidistantly. However, the log-likelihood is maximised without penalisation. Machado and Van den Hout (2018) proposed a penalised likelihood method to estimate semiparametric multi-state models with splines. The smoothing parameters are selected by using grid search. Even though the method is general and allows for backward transitions, it can become burdensome for applications that involve multiple penalties. Therefore, the methods available in the literature cannot fully address the problem of estimating multi-state models with splines for interval-censored data as they are not feasible for many applications.

In the presence of interval censoring, specific methods are needed to fit time-dependent multi-state models. For progressive processes with a limited number of states, unknown transition times can be integrated out; see, for example, Joly et al. (2002), Van den Hout (2017), and Cook and Lawless (2018). For more complex processes, particularly those with backward transitions, a piecewise-constant approximation to the time dependency can be adopted; see, for example, Kalbfleisch and Lawless (1985), Jackson (2011), and Machado and Van den Hout (2018). As mentioned above, Titman (2011) is an exception, as he fits time-dependent models to interval-censored data using direct numerical solution to the Kolmogorov Forward differential equations.

In this paper, we propose a new efficient method to estimate multi-state models with splines for interval-censored data. A Markov process framework is used to formulate the models. Hazards are specified with splines to allow for flexible modelling over time. Estimation is undertaken using a penalised likelihood approach. Given a piecewise-constant approximation to the hazards, the Fisher scoring algorithm presented in Kalbfleisch and Lawless (1985) is applied. Of specific interest is the automatic method that we present to estimate the multiple smoothing parameters. The new estimation procedure is made possible by rewriting the optimisation problem in a generalised likelihood-based framework with penalisation (Marra et al., 2017). The fitted multi-state model with splines can be used for flexible modelling of time dependency, but also to check parametric specifications.

Section 1.1 introduces the data on cardiac allograft vasculopathy (CAV). In Section 2, the hazard models with splines are defined and the likelihood function is derived. Section 3 comprises the main methodological work; it defines the penalised likelihood function and discusses how the smoothing pa-

Table 1: State table for the CAV data: number of times each pair of states was observed at successive observation times.

	To state		
From state	1	2	3
1	1314	223	136
2	0	411	105

rameters are estimated along with the model parameters. A simulation study in Section 4 shows that the proposed method works and also illustrates some effects of interval censoring. Section 5 presents the main application which is an analysis of the CAV data, and Section 6 briefly presents an additional analysis of a five-state process. Section 7 is the concluding discussion. Two appendices provide technical details additional to Section 3.

1.1. Cardiac allograft vasculopathy (CAV) data

To illustrate the methods, we analyse data for cardiac allograft vasculopathy (CAV). The data come from Papworth Hospital UK and are available in the msm package (Jackson, 2011). CAV is a narrowing of the arterial walls and the main cause of death in heart transplantation patients.

The data are a series of approximately yearly angiographic examinations of heart transplant recipients. The state at each time is a grade of CAV which can be normal, moderate or severe. Dead is the absorbing state and time of death is known within one day. The data contain 2816 rows which are grouped by 614 patients and ordered by years after transplant. Each row represents an examination and contains additional covariates. The process is biologically irreversible and of particular interest is the onset of CAV.

Diagnosis of ischaemic heart disease (IHD) and donor age are known to be major risk factors of CAV onset (Titman, 2011). In order to investigate this, three-state progressive models can be defined. The states are classified as normal (1) if the patient has not developed the disease, ill (2) if the patient has developed moderate or severe CAV and dead (3) if the patient has died, see Figure 1. Follow-up data after 15 years are not used, since after this time data are scarce which may cause identifiability problems. Titman (2011) used a similar formatting of the CAV data. Table 1 gives the number of times each pair of states was observed at successive observation times.

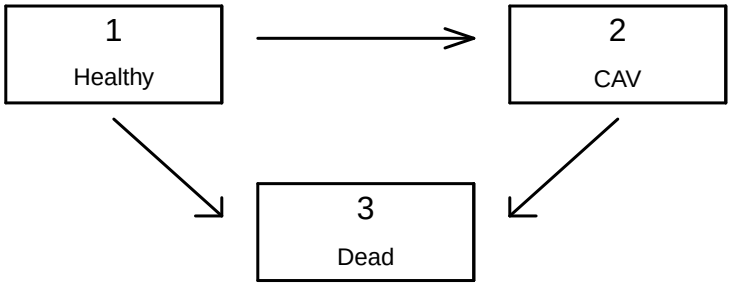

Figure 1: Illness-death model for progression of cardiac allograft vasculopathy (CAV) after transplant.

For the analysis in the current paper, we use the CAV data provided in the R package msm. This version of the data does not include the times at which patients stopped being under observation for survival follow-up. As a result, inference on time-dependent hazards may differ in certain aspects from analyses that are based on an extended version of the data; such as, for example, the analysis presented in Titman (2011).

2. Multi-state models with splines

2.1. Model representation

Let $Y(t)$ be a continuous-time Markov chain on finite state space $\mathcal{S}$, time-homogeneous transition probabilities are given by

$$p_{rs}(t_1, t_2) = P\big(Y(t_2) = s | Y(t_1) = r\big),$$

for $r, s \in \mathcal{S}$, and $t_2 \geq t_1 \geq 0$. This Markov chain is time-homogeneous because it is assumed that the probability of being in state s at time t_2 given the current state r at time t_1, depends only on the elapsed time $t_2 - t_1$. Transition matrix $\mathbf{P}(t_1, t_2)$ contains these probabilities such that the rows sum up to 1. The hazards are defined by

$$q_{rs} = \lim_{\Delta \to 0} \frac{P\big(Y(t + \Delta) = s | Y(t) = r\big)}{\Delta},$$

for $r \neq s$. The matrix with off-diagonal entries q_{rs} and diagonal entries $q_{rr} = -\sum_{r \neq s} q_{rs}$ is the generator matrix $\mathbf{Q}$. Given $\mathbf{Q}$, the solution for

$\mathbf{P}(t_1, t_2)$ subject to $\mathbf{P}(t_1, t_2) = \mathbf{I}$ for $t_2 = t_1$, is $\mathbf{P}(t_1, t_2) = \exp((t_2 - t_1)\mathbf{Q})$, see, e.g., Cox and Miller (1965).

Time-dependent models can be defined by using proportional hazards model for transition r to s, $r \neq s$,

$$q_{rs}(t) = q_{rs.0}(t) \exp\left(\boldsymbol{\beta}_{rs}^{\top}\mathbf{x}\right), \tag{1}$$

where $q_{rs.0}(t)$ is the baseline hazard function, $\mathbf{x}$ is a covariate vector and $\boldsymbol{\beta}_{rs}^{\top}$ is a vector of unknown parameters. We focus on the nonparametric estimation of $q_{rs.0}(t)$ with splines. Each hazard can be approximated by the exponential of a linear combination of K_{rs} spline basis functions $B_k(t)$ and regression coefficients $\alpha_{rs.k} \in \mathbb{R}$ as follows

$$q_{rs.0}(t) = \exp\left(\sum_{k=1}^{K_{rs}} \alpha_{rs.k} B_k(t)\right). \tag{2}$$

Let the number of spline basis functions be large (usually $K_{rs} \geq 10$) and define the vector of coefficients by $\boldsymbol{\alpha}_{rs} = (\alpha_{rs.1}, \ldots, \alpha_{rs.K_{rs}})^{\top}$ for $r \neq s$. Each $q_{rs.0}(t)$ is associated to a penalty matrix, which is quadratic in the basis coefficients and measures the complexity of $q_{rs.0}(t)$. For each transition $r \to s$, the smoothing penalty can be written as $\lambda_{rs}\boldsymbol{\alpha}_{rs}^{\top}\mathbf{S}_{rs}\boldsymbol{\alpha}_{rs}$, where $\mathbf{S}_{rs}$ is a matrix of known coefficients. The quantities λ_{rs} are called smoothing parameters and they control the trade-off between model fit and model smoothness. Large values for the smoothing parameters, $\lambda_{rs} \to \infty$, lead to a log-linear estimate of $q_{rs.0}$, while $\lambda_{rs} = 0$ results in an unpenalised regression spline estimate (Wood, 2006).

For the spline basis functions, $B_k(t)$, we use cubic regression splines which have convenient mathematical properties for multi-state modelling. However, the method is implemented in a way that is easy to employ other splines definitions and corresponding penalties.

2.2. Likelihood function

Given a multi-state model, maximum likelihood inference can be used to analyse longitudinal data. For interval-censored transition times, the likelihood function is constructed using transition probabilities. Let the state space be $\mathcal{S} = \{1, 2, .., D\}$, with D the dead state.

Let $Y_1, ..., Y_n$ be a series of states observed at times $t_1, ..., t_n$, respectively. The inference is conditional on the first observed state. For $Y_2, ..., Y_n$, the

distribution is

$$P\left(Y_n = y_n, ..., Y_2 = y_2 | Y_1 = y_1, \boldsymbol{\theta}, \mathbf{t}, \mathbf{X}\right), \tag{3}$$

where $\boldsymbol{\theta}$ is the vector with the model parameters, $\mathbf{t} = (t_1, ..., t_n)^\top$, and the $n \times p$ matrix $\mathbf{X}$ contains the values of the p covariates at each of the n time points. A conditional Markov assumption is used to define the distribution (3) as

$$\prod_{j=2}^{n} P\left(Y_j = y_j | Y_{j-1} = y_{j-1}, \boldsymbol{\theta}, t_{j-1}, \mathbf{x}_{j-1}\right),$$

where $\mathbf{x}_{j-1}$ is the $(j-1)^{th}$ row in $\mathbf{X}$. Given N individuals, the likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=2}^{n_i} P\left(Y_{ij} = y_{ij} | Y_{ij-1} = y_{ij-1}\right), \tag{4}$$

where n_i is the number of observation times for individual i.

If time of death is known, the likelihood contribution of the interval $(t_{n-1}, t_n]$ in which an individual is observed alive at time t_{n-1} and subsequently dead at time t_n is given by $\sum_{s=1}^{D-1} P\left(Y_n = s | Y_{n-1} = y_{n-1}\right) q_{sD}(t_{n-1})$. A similar definition of the likelihood can be found in Jackson (2011); see also the next section.

2.3. Piecewise-constant hazards

Time-dependency of the hazard model (1) can be taken into account by using a piecewise-constant approximation. In longitudinal data for continuous-time models, follow-up times often vary across individuals. If that is the case, the individual-specific follow-up times can be used to define the piecewise-constant approximation for the individual likelihood contributions. This implies that a transition probability such $P\left(Y_j = y_j | Y_{j-1} = y_{j-1}\right)$ is derived by using $\mathbf{Q}(t_{j-1})$ to estimate $\mathbf{P}(t_{j-1}, t_j)$ by $\exp((t_j - t_{j-1})\mathbf{Q}(t_{j-1}))$. It is also possible to impose a fixed grid to the piecewise-constant approximation as described in Van den Hout and Matthews (2008). For most applications, both methods lead to similar result and the method described in this section is preferable as it is less computationally extensive.

Using the data to define the piecewise-constant approximation explains the definition of the likelihood contribution for an observed death at time t_n in the previous section. This contribution is defined using the hazard evaluated at t_{n-1}.

3. Penalised maximum likelihood estimation

3.1. Penalised log-likelihood function

For each hazard, let the number of splines basis functions be large enough to allow for flexible modelling; see Section 2.1. This number can vary according to the time range in the data, or the number of observations. This is illustrated in the simulation study and the applications.

Let $\boldsymbol{\theta}$ be the full set of parameter and $\ell(\boldsymbol{\theta})$ be the logarithm of the likelihood function. The amount of smoothing is controlled by adding a smoothness penalty to the log-likelihood function. The penalised log-likelihood function is given by

$$\ell_p(\boldsymbol{\theta}) \;=\; \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\theta}, \tag{5}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$, vector $\boldsymbol{\lambda}$ contains the the smoothing parameters, and $\mathbf{S}_{\boldsymbol{\lambda}}$ is the penalty matrix. Matrix $\mathbf{S}_{\boldsymbol{\lambda}}$ is a block diagonal matrix with blocks $\lambda_{rs}\mathbf{S}_{rs}$ for penalising splines parameters of transition r to s and zeros elsewhere.

3.2. Parameter estimation

Given a piecewise-constant approximation to the time dependency in the hazard model (1), a scoring algorithm can be used to maximise the penalised log-likelihood function (5); see Machado and Van den Hout (2018). For a given multi-state model, if more than one hazard is specified with splines, then estimation of $\boldsymbol{\lambda}$ by direct grid search can be computationally burdensome.

There are methods available for automatic smoothing parameters estimation within the penalised likelihood framework; see Wood (2006) and Radice et al. (2016). For their method, the derivatives of the penalised log-likelihood function have to be split into the derivatives with relation to the linear predictors, and the derivatives of the linear predictor with relation to the model parameters. The direct use of their methods in multi-state models leads to large sparse matrices that are difficult to deal with.

Marra et al. (2017) developed a more general method for automatic smoothing, which uses the gradient and the Hessian (or Fisher information matrix) as a whole instead of components that make them up. The method consist of two parts. First, given a value for the smoothing parameters, we aim to find an estimate of the model parameters. Second, we use such an

estimate to find an update for the smoothing parameters. We next describe how to perform the first part of the method.

Let $\mathbf{g}_p^{[a]} = \mathbf{g}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\theta}^{[a]}$ and $\boldsymbol{\mathcal{I}}_p^{[a]} = \boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\boldsymbol{\lambda}}$ represent the penalised gradient and negative of the penalised hessian matrix at iteration a, respectively, where $\mathbf{g}^{[a]} = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[a]}}$ and $\boldsymbol{\mathcal{I}}^{[a]} = -\partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[a]}}$. For fixed value of $\widehat{\boldsymbol{\lambda}}$, the a^{th} estimate of $\boldsymbol{\theta}$ can be updated by

$$\boldsymbol{\theta}^{[a+1]} = \left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\right)^{-1} \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\mathbf{z}^{[a]}, \tag{6}$$

where $\mathbf{z}^{[a]} = \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\boldsymbol{\theta}^{[a]} + \boldsymbol{\epsilon}^{[a]}$ and $\boldsymbol{\epsilon}^{[a]} = \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}^{-1}\mathbf{g}^{[a]}$.

This parametrisation allows for a well founded formulation of the smoothing parameters selection presented in Section 3.3 (Marra et al., 2017); see Appendix A for a justification for this parametrisation. Calculating the second derivatives of the probability matrix can be intractable; see Kalbfleisch and Lawless (1985). We use an approximation to the Fisher information matrix that involves only the first order derivatives of the penalised log-likelihood function; see Appendix B.

3.3. Smoothing parameters estimation

The penalised maximum likelihood approach described in Section 3.2 can only estimate model parameters, $\boldsymbol{\theta}$, for fixed vector of smoothing parameters, $\boldsymbol{\lambda}$. In general, if there are only one or two smoothing parameters, a common approach to estimate these parameters is to undertake a grid search over possible values and use AIC (cf. Machado and Van den Hout 2018) or generalised cross-validation (cf. Eilers and Marx 1996; Wu and Sickles 2018). However, in our multi-state model there are multiple penalty parameters and a grid search is not feasible. In this section, we briefly discuss the automatic estimation of the smoothing parameters as presented in Marra et al. (2017).

From likelihood theory, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix, $\boldsymbol{\mu}_{\mathbf{z}} = \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\theta}$ and $\boldsymbol{\theta}$ is the true parameter vector. The predicted value vector for $\mathbf{z}$ is $\widehat{\boldsymbol{\mu}}_{\mathbf{z}} = \sqrt{\boldsymbol{\mathcal{I}}}\widehat{\boldsymbol{\theta}} = \mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{z}$, where $\mathbf{A}_{\widehat{\boldsymbol{\lambda}}} = \sqrt{\boldsymbol{\mathcal{I}}}(\boldsymbol{\mathcal{I}} + \mathbf{S}_{\widehat{\boldsymbol{\lambda}}})^{-1}\sqrt{\boldsymbol{\mathcal{I}}}$. The smoothing parameter vector is estimated to minimise

$$\mathbb{E}(||\boldsymbol{\mu}_{\mathbf{z}} - \widehat{\boldsymbol{\mu}_{\mathbf{z}}}||^2) = \mathbb{E}(||\mathbf{z} - \mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{z}||^2) - c + 2tr(\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}),$$

where c is a constant. In practice, $\boldsymbol{\lambda}$ is estimated by minimising the Un-Biased Risk Estimator (UBRE; Craven and Wahba, 1979)

$$\mathcal{V}(\boldsymbol{\lambda}) = ||\mathbf{z} - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}||^2 - c + 2tr(\mathbf{A}_{\boldsymbol{\lambda}}). \tag{7}$$

Equation (7) can be minimised using the automatic smoothing parameters selection method developed by Wood (2004) or in principle by using a general-purpose optimiser.

Hazard models defined with splines can capture parametric hazards. For an hazard that is exponential or Gompertz, the corresponding smoothing parameter λ is infinity. For this reason, it is recommended to impose an upper bound when estimating smoothing parameters; this will be illustrated in the simulation study.

3.4. Summary of the algorithm

The methods described in Sections 3.2 and 3.3 can be used to define an algorithm that iterates until the parameter estimator satisfies $\max |\boldsymbol{\theta}^{[a+1]} - \boldsymbol{\theta}^{[a]}| < \delta$ for a suitable small positive value (Radice et al., 2016). The two steps of the algorithm are as follow:

Step 1: For fixed smoothing parameters $\boldsymbol{\lambda}^{[a]}$, find an estimate of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{[a+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \ell_p(\boldsymbol{\theta}).$$

Step 2: Given the estimate $\boldsymbol{\theta}^{[a+1]}$, find an estimate of $\boldsymbol{\lambda}$ using (7):

$$\boldsymbol{\lambda}^{[a+1]} = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}}\ \mathcal{V}(\boldsymbol{\lambda}).$$

3.5. Confidence intervals

The distribution of the penalised maximum likelihood estimator can be used to construct confidence intervals for non-linear functions of the estimate $\widehat{\boldsymbol{\theta}}$, such as the hazards and probability matrices (Wood, 2006). Let $\mathbf{V}_{\boldsymbol{\theta}}$ represent the covariance matrix of $\widehat{\boldsymbol{\theta}}$ at convergence. From large sample theory, samples of the estimate $\widehat{\boldsymbol{\theta}}$ can be drawn from $N(\widehat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$. Confidence intervals for functions of the model parameters can be constructed as follows:

Step 1: Draw b vectors from $N(\widehat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$.

Step 2: Calculate the value of the function of interest at each simulated value.

Step 3: Using the simulated values of the function, calculate the lower $(\varsigma/2)$ and upper $(1-\varsigma)$, quantiles.

The parameter ς is usually set to 0.05. In this paper, we approximate the covariance matrix $\mathbf{V}_{\boldsymbol{\theta}}$ by the inverse of the matrix $\mathbf{M}$ described in Appendix B.

4. Simulation study

We perform a simulation study to investigate the performance of the method presented in Section 3 for modelling time-dependency in multi-state processes. The study will investigate the fitting of various hazard shapes, the effect of interval censoring, and a comparison with other methods. The simulation is for a progressive three-state illness-death process as shown in Figure 1. For each of the three hazards we will assume a different time-dependent shape so that we can show that the model with splines can deal with various scenarios. The time dependency will be simulated using parametric distributions, but the fitted hazards will be based on three-state models with splines. The simulation study is implemented in the R software; the code can be obtained by contacting the authors.

4.1. Scenarios

For a three-state progressive process, we define the transition hazard for $1 \rightarrow 2$ using a log-normal distribution with parameters $\mu = 1.25$ and $\sigma = 1$. This implies that the hazard increases at first and decreases at a later time. We define a constant hazard for $1 \rightarrow 3$ using an exponential distribution with rate $\exp(-2.5)$. The hazard for $2 \rightarrow 3$ is defined to increase strictly by using the Gompertz distribution with rate $\exp(-2.5)$ and shape 0.1.

Given the above parametric assumptions, longitudinal data are simulated repeatedly for N individuals. An individual illness-death trajectory is simulated in two steps. First, transition times are simulated. Next a longitudinal sample design is imposed so that transition times to state 2 are interval censored.

Using years since baseline as the time scale, transition times are simulated as follows. Let $T_{rs} = T_{rs|u}$ represent the time of the transition to state s conditional on being in state r at time $u > 0$. If state at u is 1, then the time of transition to the next state can be obtained by taking $T = \min\{T_{12}, T_{13}\}$. If $T = T_{12}$ then, the next state is 2, otherwise the next state is 3. If state is 2, then the time of the next state is T_{23}. The event times T_{12} are simulated using the function rgengamma in R (Jackson, 2016). The event times T_{13}

and T_{23} are simulated with user-written code for the exponential and the Gompertz distribution.

A sampling design is imposed by assuming that living states are observed at years $t_1 = 0, t_2, t_3, ..., , t_n = 15$, where the times are in years. This leads to interval-censored transition times for transitions $1 \to 2$. Death times that are simulated within the 15-year period are used as exact times for the transitions to the dead state. We will investigate several sequences of $t_1, t_2, ..., t_n$.

Given the simulated data, the hazards functions are estimated using splines. The package mgcv (Wood, 2007) in R is used to set the design and penalty matrices. The number of spline basis functions (number of knots) for each hazard is K, hence the model has a total of $3K$ parameters. We use cubic regression splines, in which case the knots are placed using the percentiles of the observation times. Therefore, knots placement is different for every sample. The three-state model with splines is then estimated using the procedure described in Section 3. The smoothing parameters are estimated using the general-purpose optimiser optim in R. To prevent numerical problems with smoothing parameters $\lambda = \exp(\gamma)$ estimated at infinity for exponential and Gompertz shapes, we impose upper bound $\gamma < 20$; see also Section 3.3.

4.2. Numerical results for 100 replications

The scenario above is repeated 100 times for $N = 500$ individuals, times $(t_1, t_2, t_3, ..., t_n) = (1, 2, 3, ..., 15)$, and $K = 10$. Figure 2 presents the comparison between estimated hazards and true hazards. The black lines represents the true hazards and the white lines the medians of the estimated hazards.

The means of the estimated hazards (not shown) are quite similar to the medians up to around 14 years, with some overestimation in the last year for hazards $1 \to 2$ and $1 \to 3$. The medians are robust to those few hazards that were fitted with relatively high values at the later years. This large variation between the fitted hazards towards the end of study time (as illustrated by the grey curves in Figure 2) is due to scarceness of data at later years.

Figure 2 shows that the method seems to work quite well overall. Nevertheless, there is some discrepancy between the true hazard for $1 \to 2$ and the median. The fact that the true hazard starts at zero is not represented accurately in the fitted hazards. This is due to the interval censoring defined in the simulation. The sampling design is such that the living states are observed at intervals of one year. For the first two years after baseline, this design does not work well. We investigate this further in the next section.

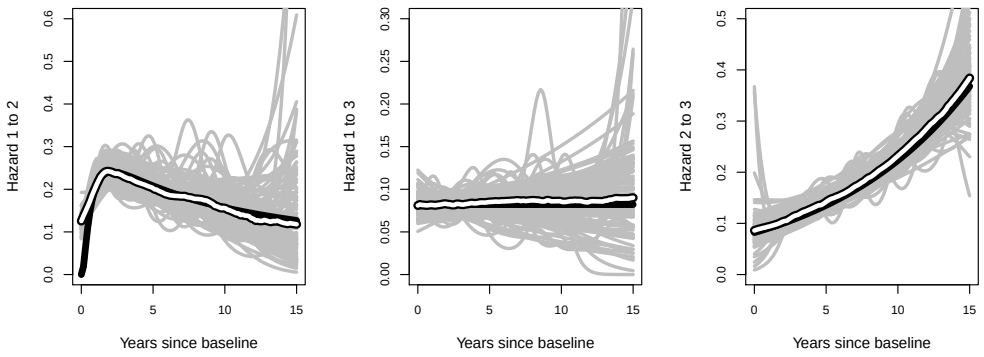

Figure 2: Simulation study for the illness-death model using a yearly follow-up with $N = 500$ individuals. True hazards (black lines), estimated hazards (grey lines, 100 replications), and the medians of the estimated hazards (white lines).

Table 2: Simulation study to investigate the performance of the multi-state models with splines for modelling time-dependent processes. For 10-year transition probabilities, mean, median, and bias for $R = 100$ replications.

Probabilities	True	Mean	and bias	Median	and bias
$p_{11}(0, 10)$	0.065	0.060	0.004	0.060	0.005
$p_{12}(0, 10)$	0.231	0.222	0.009	0.222	0.007
$p_{13}(0, 10)$	0.704	0.718	-0.014	0.718	-0.012
$p_{22}(0, 10)$	0.245	0.231	0.014	0.231	0.016
$p_{23}(0, 10)$	0.755	0.769	-0.014	0.769	-0.016

Data are simulated using parametric hazards, but models are fitted using splines. Hence, we cannot compare true parameters values with estimated parameter values in the current simulation study. However, we can compare summary statistics computed using the true parametric hazards with statistics computed from the fitted splines. An example of such a statistic is the transition probability matrix for a given time interval.

Table 2 presents the results of the simulation study in terms of transition probabilities. It shows the ten-year transition probabilities for the true model, the median and mean of the estimated ten-year transition probabilities, and the corresponding bias. The results show that the multi-state model with splines can estimate transition probabilities well for the ten-year time interval $(0, 10]$.

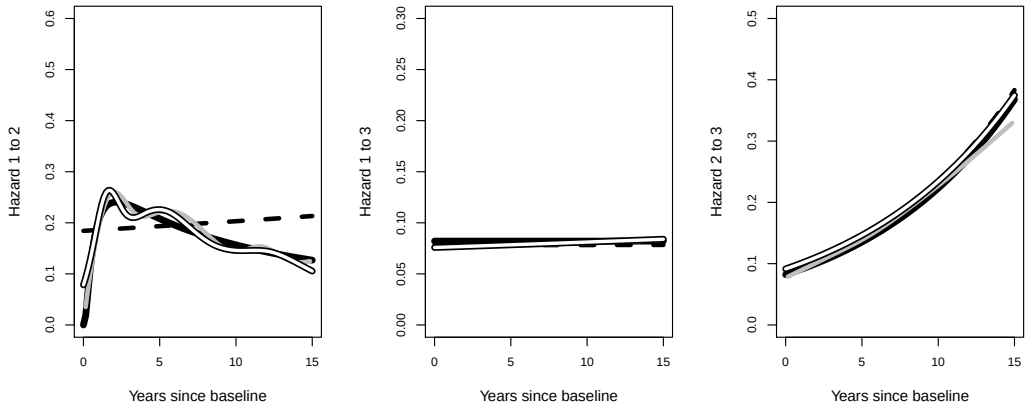

Figure 3: One-off replication for the illness-death model using $N = 2500$ individuals. True hazards (black lines), hazards estimated using msm (dashed black lines), using SmoothHazard (grey lines), and using the proposed method (white lines).

4.3. Comparison with other methods

To compare our method with other methods, and to further investigate the effect of interval censoring we present a simulation with one replication only.

The scenario in Section 4.1 is adapted by adding planned observation times in the first two years. The design for the follow-up times $(t_1, t_2, t_3, ..., t_n)$ in years is $(0, 0.5, 1, 1.5, 2, 3, 4, ..., 15)$. As before, we use $K = 10$. We simulated this scenario once for $N = 2500$ individuals, and fit a parametric time-dependent model and two spline models.

The parametric model is defined using the Gompertz hazard in (1); that is, by defining $q_{rs}(t) = q_{rs.0}(t) = \exp(\beta_r s + \xi_{rs}(t))$. The model is fitted using the package msm (Jackson, 2011).

The first spline model is fitted using the function idm in the package SmoothHazard (Joly et al., 2002). We specify 10 knots for each transition, and use the option in the software to estimate the three smoothing parameters by approximated cross validation. The second spline model is fitted using our method as specified at the end of Section 4.1.

Figure 3 compares the true hazard with the fitting of the three models. We see that the Gompertz model fits well for the hazards for $1 \to 3$ and $2 \to 3$ despite model misspecification of the hazard for $1 \to 2$.

All three true hazards are captured quite well by the model fitted with

idm and the spline model fitted with our algorithm. The former captures the hazard for $1 \to 3$ a bit better at the start of the time scale, but the latter is better at smoothing the hazard for $2 \to 3$ at later years.

There is quite a difference in computation time needed for the three models above. Using a laptop with Windows 7 (2GHz processor, 8GB RAM, 64-bit), the Gompertz is fitted by msm very fast (3.03 seconds). Fitting models with cross-validation in idm is computationally intensive. Using the default settings, the above model needed 100.3 minutes.

We implemented our algorithm in R without using parallel computing or internal routines in other programming languages. The algorithm for the spline model was relatively fast, it took 15.7 minutes. The initial values for the spline weights were $\alpha_{rs} = -3$ for the relevant (r, s), and penalty vector $\boldsymbol{\lambda} = \exp(\boldsymbol{\gamma})$ had initial values $\boldsymbol{\gamma} = (1, 1, 1)$. Estimated $\boldsymbol{\gamma}$ is $(1.76, 19.97, 19.99)$. The estimated penalties for $1 \to 3$ and $2 \to 3$ are at the imposed upper bound for the entries in $\boldsymbol{\gamma}$, which is in agreement with the loglinear shape of the simulated hazards for these transitions.

4.4. Interval censoring and identifiability

Substantial interval censoring can have a deteriorating impact on estimation. Consider the sample design in the current simulation design: observations of state 1 and 2 are restricted to pre-specified times $(t_1, t_2, t_3, ..., t_n)$. If the time between observations is this sample design is large relative to the change in the hazards, this may lead to an identifiability problem. Following a suggestion from an anonymous reviewer of this paper, we investigate this by simulating data given a very specific shape of the hazard for transition $1 \to 2$.

In Figure 4, the black line is the true cosine-shape hazard from which we simulated transitions from state 1 to 2. For the other transitions, we use the same parametric shapes as before: exponential for $1 \to 2$, and Gompertz for $2 \to 3$. The reason for using a cosine-shape hazard is as follows: if we only have observation of state 1 and 2 at times $(t_1, t_2, t_3, t_4) = (0, 5, 10, 15)$, then we expect to be able to identify possible change of the hazard for $1 \to 2$ from time t_j to t_{j+1}, for $j = 1, 2, 3$, but may fail to identify the hazard in more detail. Given the cosine-shape hazard, this might imply that we are not able to distinguish a fitted sine shape from the true cosine shape.

We fit a three-state process for $N = 500$ individuals and apply two designs for interval censoring. Design $\mathcal{A}$ is defined by $(t_1, t_2, t_3, ..., t_n) =$

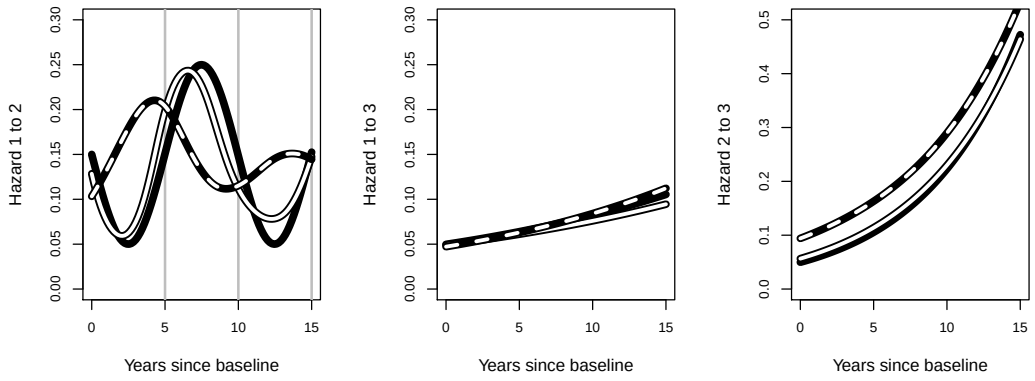

Figure 4: One-off replication for the illness-death model using $N = 500$ individuals. True hazards (black lines), hazards estimated using the spline model with $K = 10$ and one-year follow-up data (white lines), and hazards estimated using the spline model with $K = 5$ and five-year follow-up data (dashed white lines).

$(1, 2, 3, ..., 15)$. Using $K = 10$, the spline model recovers the true hazards reasonably as illustrated by the white line in Figure 4.

Next we define design $\mathcal{B}$ by $(t_1, t_2, t_3, t_4) = (0, 5, 10, 15)$. In this case, we have the same underlying process as with $\mathcal{A}$, but the follow-up information is more sparse. At the times (t_1, t_2, t_3, t_4), however, the distribution of observed states is the same for $\mathcal{A}$ and $\mathcal{B}$.

For the data given design $\mathcal{B}$, we could not fit a spline model with $K = 10$, but for $K = 5$ the algorithm converges without problems. And the identifiability problem is nicely illustrated by the bad fit in Figure 4: instead of a cosine shape, a sine shape is fitted for the hazard for $1 \to 2$. This misfit also leads to a bad fit for the hazard for $2 \to 3$. Figure 4 also shows that the identification at the observation times (t_1, t_2, t_3, t_4) is decent. In this case, we are able to more or less identify the change of the hazard for $1 \to 2$ from time t_j to t_{j+1}, for $j = 1, 2, 3$, but fail to identify this hazard within the corresponding time intervals $(t_j, t_{j+1}]$.

Given identifiability problems, estimation is often sensitive to starting values. We explored this also in current case, but ended up with a sine shape for other starting values as well.

4.5. Conclusion simulation study

The findings from the simulation results are threefold. First, they indicate that the proposed method is able to estimate nonlinear, log-linear and linear hazards in the presence of interval censoring. Second, they show that the piecewise-constant approximation to the transition probabilities provides satisfactory results, as we are able to recover the true curves and ten-year transition probabilities. Third, although interval censoring can be dealt with, the results show that it can also lead to problems if the censoring is substantial relative to the volatility of the underlying process.

5. Application to CAV data

We fit an illness-death model for the CAV data defined as in Figure 1. Because time of death is known within one day, rather than being interval censored, the likelihood contribution of individuals observed in state $r < 3$ at time t and dead at time $t^* > t$ are given by $\sum_{s=1}^{2} P\left(Y(t^*) = s | Y(t) = r\right) q_{s3}(t)$. As described in Section 2.3, transition probabilities for the likelihood function are calculated by using a piecewise-constant approximation to the hazards. For the CAV data, the mean length of the interval between observations within one patient is 1.622 years with standard deviation of 0.972 and median 1.258. Assuming that change of health status can be assessed in intervals of approximately 1.2 years, we can use the data to define the grid for the piecewise-constant approximation.

Let t represent time since transplant. The proportional hazard model with splines is specified with dependence on donor age ($dage$) and primary diagnosis of ischaemic heart disease (IHD):

$$q_{rs}(t) = \exp\left(\sum_{k=1}^{10} \alpha_{rs.k} B_k(t) + \beta_1 dage + \beta_2 IHD\right), \tag{8}$$

where $(r, s) \in \{(1,2),(1,3),(2,3)\}$ and $B_k(t)$ are known spline basis functions. We use penalised cubic regression splines. The knots are placed considering the percentiles of the observation times. This is a key factor for fitting multi-state models with splines. Because multi-state data can become scarce close to the end of study, there might not be enough information to estimate some basis coefficients. Fitting multi-state models with P-splines (Eilers and Marx, 1996) might not be possible for some applications as it requires the knots to be equally spaced. In that case some knots might be

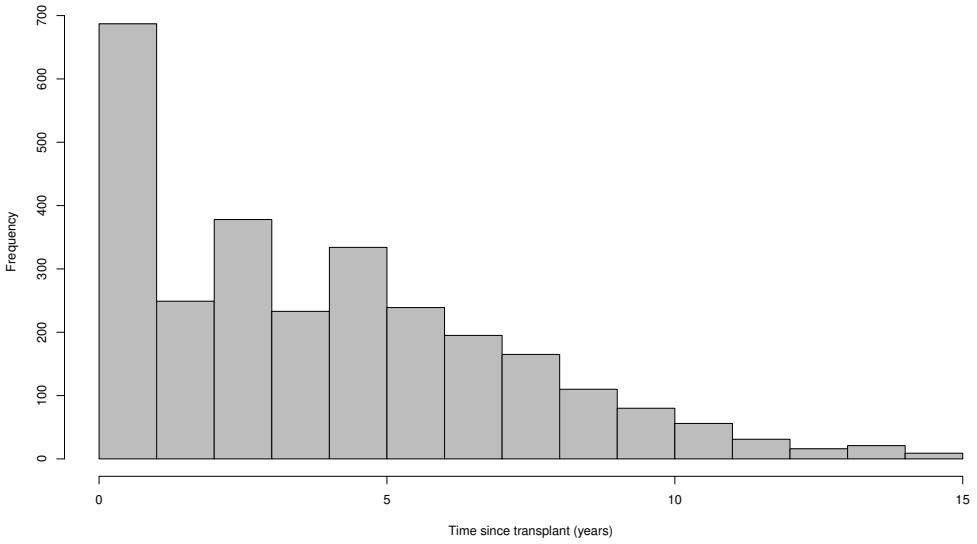

Figure 5: Histogram of time since transplant in the CAV data.

placed where there is no data. Figure 5 illustrates the histogram of time since transplant for the CAV data.

For the analysis to follow, the design and penalty matrices are set up using the package mgcv in R. As indicated in (8), the hazards are modelled with 10 knots each, hence the total number of parameters is 32. The vector of smoothing parameters is $\boldsymbol{\lambda}^\top = (\lambda_{12}, \lambda_{13}, \lambda_{23})$. The multi-state model with splines is estimated using the procedure described in Section 3. The smoothing parameters are estimated using the general-purpose optimiser optim in R.

The estimated smooth hazards for subjects with IHD and donor age of 26 (solid lines) and 95% confidence intervals (dashed lines) are presented in Figure 6. The risk of moving from state 1 (healthy) to state 2 (CAV) increases until approximately 8 years after transplant, but decreases afterwards. The risk of going from state 1 to state 3 (dead) is very low and almost constant until approximately 10 years since transplant, but increases pretty steep afterwards. The transition intensity from state 2 to state 3 is quite volatile and upwards until 10 years after transplant and decreasing afterwards. The confidence intervals are fairly wide after approximately 10 years, which is to be expected given that data become scarce after 10 years.

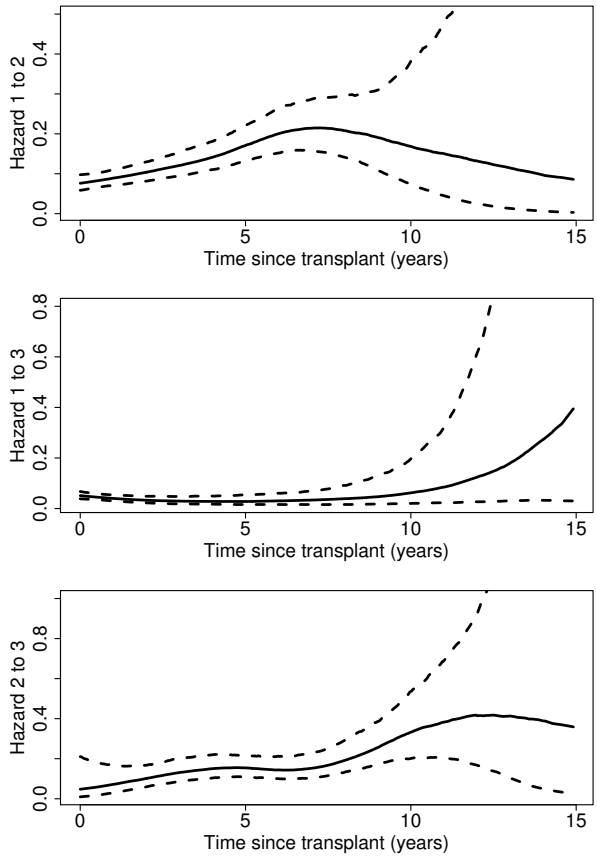

Figure 6: Estimated smooth hazards for subjects with IHD and with donor age of 26 (solid lines), with 95% confidence intervals (dashed lines).

For the parametric part of the model, $\widehat{\beta_1} = 0.018$ (0.04) and $\widehat{\beta_2} = 0.274$ (0.096) indicating that donor age and IHD increase the risks of disease progression and death. The vector of smoothing parameters is estimated at $\widehat{\boldsymbol{\lambda}} = (47.145, 41.668, 10.716)^\top$.

Although estimated hazards gives insightful information about the risks of moving across states, interpretation is more straightforward when transition probabilities are considered. For subject with IHD and with donor age of 26, the five-year transition probabilities are estimated at

$$\widehat{\mathbf{P}}(0,5) = \begin{pmatrix} 0.475\ (0.412, 0.529) & 0.291\ (0.250, 0.335) & 0.234\ (0.197, 0.286) \\ 0 & 0.579\ (0.428, 0.675) & 0.421\ (0.325, 0.572) \\ 0 & 0 & 1 \end{pmatrix}, \quad (9)$$

with 95% confidence interval (in brackets) obtained using $b = 1000$ simulations as in Section 3.5. A transition probability can be interpreted as follows. A subject with IHD and donor age of 26 has a 29% chance of being in the CAV five years later.

To further illustrate the improvement achieved by the method presented in this paper, we compare the fit of the model with splines (8) with the fit of a model with Gompertz hazards specification. The Gompertz hazards specification is common in parametric multi-state modelling due to its simplicity and straightforward use within the msm package; see, for example, Robitaille et al. (2018) and Marioni et al. (2012).

The proportional hazard model with Gompertz specified with dependence on donor age ($dage$) and primary diagnosis of ischaemic heart disease (IHD) is given by

$$q_{rs}(t) = \exp\left(\alpha_{rs} + \xi_{rs}t + \beta_1^* dage + \beta_2^* IHD\right), \tag{10}$$

where $(r, s) \in \{(1, 2), (1, 3), (2, 3)\}$. The model is estimated using a scoring algorithm. The covariates effects and standard errors (in brackets) are estimated at $\widehat{\beta_1^*} = 0.018$ (0.04) and $\widehat{\beta_2^*} = 0.277$ (0.094). Then the covariates effects and their standard errors for models (8) and (10) are equivalent.

Model validation for multi-state models can be carried out by comparing model prediction of the entry time into the dead state with the Kaplan-Meier curve estimates (Titman and Sharples, 2010). Figure 7 depicts baseline-specific survival as estimated by the models (10) and (8) (on the left and right hand side, respectively) and as described by the Kaplan-Meier curves. For the Gompertz model in (10), the fit is reasonably good up to 10 years, but after that the model fails to predict survival. The multi-state model with splines predicts the survival reasonably accurately throughout the years.

6. Application to ELSA data

To illustrate our method with a five-state process, we analyse data from the English Longitudinal Study of Ageing (ELSA, `www.ifs.org.uk/ELSA`). The ELSA baseline is a representative sample of the English population aged 50 and older. ELSA contains information on health, economic position, and quality of life. Here we use a random sample of 1,000 individuals, with 544 women and 456 men. The number of observations per individual ranges from 2 up to 6. This is the same sample as used in Van den Hout (2017). ELSA data are made available through the UK Economic and Social Data Service

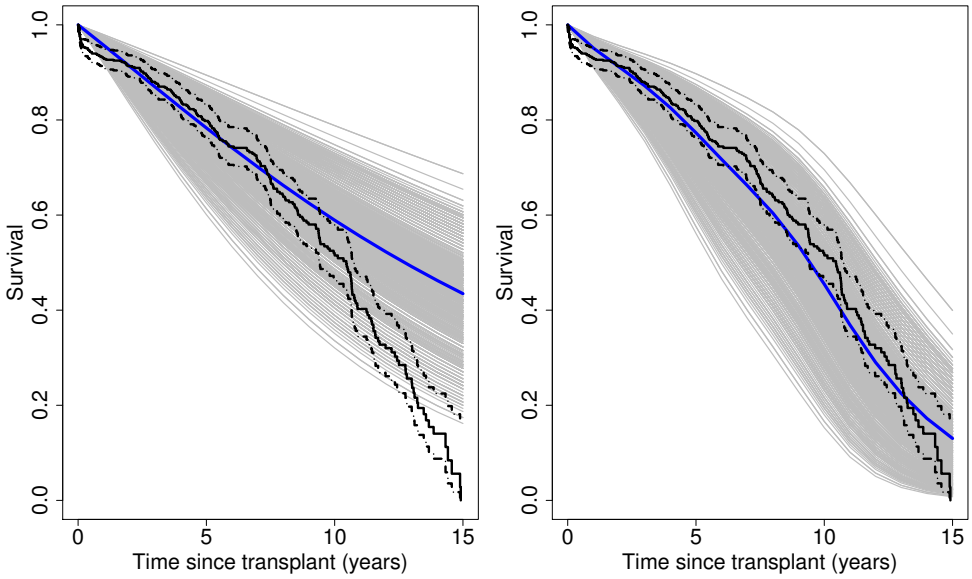

Figure 7: Comparison of model-based survival with Kaplan-Meier curves for the Gompertz model (left-hand side) and spline model (right-hand side). Model-based survival: grey lines for individuals, blue lines for the mean of the individual curves. Kaplan-Meier in black lines with 95% confidence intervals.

Table 3: State table for the ELSA data: number of times each pair of states was observed at successive observation times. The four living states are defined by number of words remembered.

	To				
From	10-7 words	6-5 words	4-2 words	1-0 words	Dead
10-7 words	164	150	49	12	8
6-5 words	156	440	303	48	40
4-2 words	52	336	616	151	85
1-0 words	11	35	114	149	72

(`www.esds.ac.uk`). In the data that we use, information on age is rounded to integers for reasons of data protection.

In this application, we define four living states using the score on a word-recall test. During the interview, individuals are asked to remember words from a list of 10 that was read out aloud at an earlier time in the same interview. The living states are defined by the number of words an individual can remember: state 1, 2, 3, and 4, for number of words $\{7, 8, 9, 10\}$, $\{6, 5\}$, $\{4, 3, 2\}$, and $\{1, 0\}$, respectively. We define the fifth state as the dead state.

The interval-censored five-state data are summarised by the frequencies in Table 3. The sum of the transitions into the dead state is equal to the number of deaths in the sample; that is, 205. Table 3 shows that the process includes backwards transitions between the living states.

We define three models to illustrate the spline modelling. The intercepts-only model is given by

$$q_{rs}(t) = \exp(\beta_{rs}) \qquad \text{for} \qquad (r, s) \in \big\{(1,2), (1,5), (2,1), (2,3), (2,5) \\ (3,2), (3,4), (3,5), (4,3), (4,5)\big\}.$$

In our models, we assume that transitions between two states that are not contiguous imply visiting the intermediate state(s) at least once. For example, direct transitions from state 1 to 3 are not possible—the process has to go via state 2. With 10 possible transitions, the intercepts-only model has 10 parameters. The AIC is 8109.5.

Given that change of cognition is likely to be associated with changing

age, we use age as time scale t in the second model given by

$$q_{rs}(t) = \exp(\beta_{rs} + \xi_{rs} t) \qquad \text{where } \xi_{rs} = 0 \text{ for } (r,s) \in \big\{(1,5),(2,1),(3,2),(4,3)\big\} \; . \quad (11)$$

In this model, intercept β_{rs} is included for all the 10 possible transitions. The justification of the restrictions on the age effects is twofold. Table 3 shows that there is not a lot of information on transition $1 \to 5$, hence $\xi_{15} = 0$. The other restrictions are imposed to define a parsimonious model. Interest in cognitive change is often focussed on decline, hence we keep the model simple for the backward transitions in our process. This model has 16 parameters, and the AIC is 7962.6. The improvement in the AIC illustrates the importance of taking age into account.

In model (11), the parametric shape is defined by the Gompertz distribution. This is quite restrictive. We define a more flexible spline model by using 5 knots for each transition for which we defined an age effect in model (11). We use the same approach as with the CAV albeit for an extended number of transitions: penalised cubic regression splines are used, with knots being placed using the percentiles of the observation times. The spline model has 10 intercepts and 6×5 spline weights, hence 40 model parameters in total. For each of the 6 fitted splines, we have a separate penalty parameter $\lambda = \exp(\gamma)$ which is estimated using upper bound $\gamma < 20$. The AIC for this model is 7800.4, which is based on an effective number parameters equal to 19.4.

Figure 8 shows the fitted hazards for the parametric model and the spline model. For the transitions with the constant hazard, fitted hazards are similar. For the other transitions, we see some discrepancy between fitted parametric curves and the splines. Given the reduction in the AIC when using the spline model compared to the parametric model (11), we infer that the former describes the transitions hazards better then the latter.

The fitted splines in Figure 8 are quite smooth. We think that this is due to the substantial interval censoring in the ELSA data. The median of the time intervals between interviews within one individual is 2 years. This implies that information on the time of changing state is limited across the age range in the data. Hence the smoothness of the fitted hazards.

The limited information across the age range is probably also the reason that we were not able to fit spline models with $K = 10$ knots for the relevant hazards.

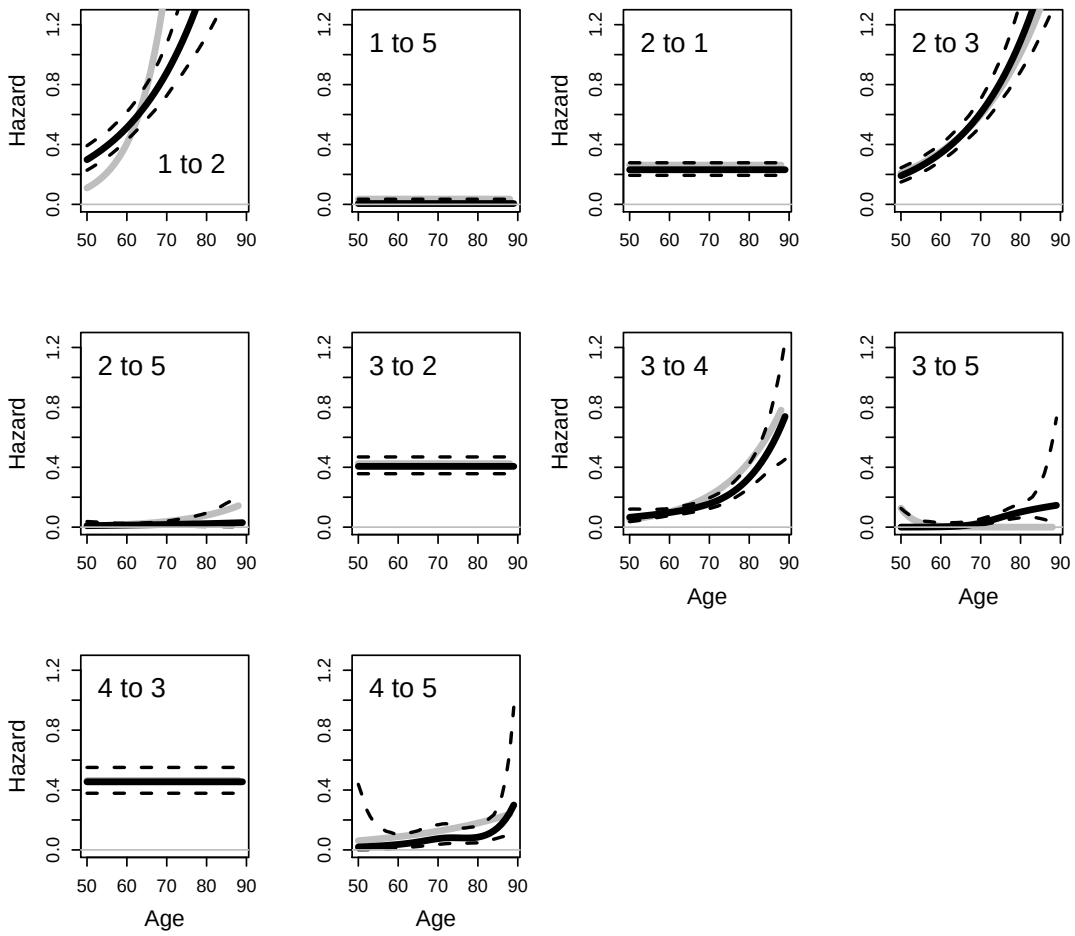

Figure 8: Estimated hazards in the five-state model for the ELSA data. Grey lines for the parametric model, and black lines for the spline model (with dashed lines for the 95% confidence intervals).

7. Discussion

This paper presents a practical and unifying framework for estimating multi-state models with splines for interval-censored data. The new estimation method is made possible by rewriting the optimisation problem using a penalised general likelihood estimation (Marra et al., 2017).

The simulation study shows the importance of the method for flexible modelling of time-dependent processes. It is shown that the method can recover nonlinear, log-linear and linear hazards in the presence of interval censoring. However, the simulation study also illustrates that estimation can be biased if the interval censoring is substantial relative to the volatility of the underlying process.

The method is applied to a three-state illness-death process without recovery for cardiac allograft vasculopathy (CAV), and to a five-state process for cognitive function and mortality in the English Longitudinal Survey of Ageing (ELSA). These applications illustrate the feasibility of the method and its usage for flexible time-dependent modelling. There should not be a problem to apply the method for more complex multi-state processes, as long as there are enough observations for those transitions that are modelled with splines. Another application can be found in Machado (2018), where a four-state model with one backward transition is investigated for an ageing process. In that example, the choice of Gompertz distributions is justified by showing that the penalised-splines hazards closely resemble the parameteric Gompertz hazards.

The simulation study and the applications also show potential problems when our method is applied in practice. Scarceness of data can be an issue. If there is a dead state, then multi-state data become scarcer during the study follow-up. Although our method can be implemented for various type of splines, splines where the placing of the knots is dependent on observation times are to be preferred when data become scarcer during the follow-up. When data are scarce or when there is substantial interval censoring, the fitted shapes of the transition hazards should be interpreted with care. The simulation study illustrated potential bias in a rather extreme case. The analysis of the ELSA data shows a more subtle situation where fitted hazards are very smooth—a feature that may be partly due to limited information on changing state across the age range in the data.

The automatic smoothing parameters estimation as described in Marra et al. (2017) requires the Hessian or the Fisher information for estimation.

With the simulation study an the applications, we show that an approximation to the Fisher information matrix, which only uses the first order derivatives of the log-likelihood, performs well on estimation. This is relevant for interval-censored data as calculating the second derivatives of the transition probabilities can be intractable.

As discussed in Titman (2011), CAV is a progressive disease even though backward transitions are recorded, due to measurement errors. The work presented here can be extended to allow for misclassification of states (Jackson et al., 2003). A similar extension might be useful for the analysis of the ELSA data in Section 6, where the definition of the states by the number of words recalled is likely to be subject to test-retest error. Allowing for misclassification of state in the spline models poses extra difficulty for estimation as derivative free algorithms, e.g., a quasi-Newton algorithm, are required to maximise the penalised log-likelihood function.

This paper shows how penalised splines can be used to model time-dependent transition hazards. The same method can also be used to deal with time-dependent covariates in regression models for the hazards. As an example, one might be interested in the effect of changing blood pressure on disease onset. In such a case, interval censoring needs extra attention. It is possible that time intervals between observations are informative with respect to change of disease but are too long to provide good information about a time-dependent covariate.

Penalised splines can also be used in multi-state models outside the framework in the current paper. For example, models that take time spent in a state into account, or models that deal with left-truncation in observational studies.

The msm package (Jackson, 2011) was primarily designed to model time-homogeneous multi-state models. However, it is possible to fit some time-dependent models, such as Gompertz and splines (without penalties) models. In this case, time-dependency is dealt with by msm by using a piecewise-constant approximation to the hazards. Our method used the same piecewise-constant approximation and can thus be seen as a generalisation of the method in msm.

The Gompertz hazards specification is common in many applications due to its simplicity and straightforward use with the msm package. We show through a model validation method that such restrictive model specifications can lead to poor model fit. As shown in Figure 7, the multi-state model with splines can improve considerably model fit by allowing for flexible hazards

specification.

Acknowledgments

This research was supported by CNPq - Brazil [249308/2013-4].

The ELSA data were made available through the UK Economic and Social Data Service. ELSA was developed by a team of researchers based at the National Centre for Social Research, University College London and the Institute for Fiscal Studies. The developers and funders of ELSA do not bear any responsibility for the analyses or interpretations presented here.

The authors would like to thank the Associate Editor and two reviewers for substantial feedback that has helped to improve the manuscript.

Appendix A. Parametrisation in the estimation

For easy reference, we derive the parametrisation of the model-parameter estimators as in Marra et al. (2017). A first-order Taylor expansion of $\mathbf{g}_p^{[a+1]}$ about the current fit $\boldsymbol{\theta}^{[a]}$ is given by

$$\mathbf{g}_p^{[a+1]} \approx \mathbf{g}_p^{[a]} + \boldsymbol{\mathcal{H}}_p^{[a]}(\boldsymbol{\theta}^{[a+1]} - \boldsymbol{\theta}^{[a]}), \tag{A.1}$$

where $\mathbf{g}_p^{[a]} = \mathbf{g}^{[a]} - \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\theta}^{[a]}$ and $\boldsymbol{\mathcal{H}}_p^{[a]} = \boldsymbol{\mathcal{H}}^{[a]} - \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}$. Let us define $\boldsymbol{\mathcal{I}}^{[a]} = -\boldsymbol{\mathcal{H}}^{[a]}$. A new fit $\boldsymbol{\theta}^{[a+1]}$ is obtained by taking the right-hand side of equation (A.1) to be zero

$$\begin{aligned}
\mathbf{0} &= \mathbf{g}_p^{[a]} + \left(-\boldsymbol{\mathcal{I}}^{[a]} - \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\right)(\boldsymbol{\theta}^{[a+1]} - \boldsymbol{\theta}^{[a]}) \,, \\
\mathbf{g}_p^{[a]} &= \left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\right)(\boldsymbol{\theta}^{[a+1]} - \boldsymbol{\theta}^{[a]}) \,, \\
\mathbf{g}^{[a]} - \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\theta}^{[a]} &= \left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\right)\boldsymbol{\theta}^{[a+1]} - \boldsymbol{\mathcal{I}}^{[a]}\boldsymbol{\theta}^{[a]} - \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\theta}^{[a]} \,, \\
\left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\right)\boldsymbol{\theta}^{[a+1]} &= \mathbf{g}^{[a]} + \boldsymbol{\mathcal{I}}^{[a]}\boldsymbol{\theta}^{[a]} \quad \text{and} \\
\boldsymbol{\theta}^{[a+1]} &= \left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\left(\sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\boldsymbol{\theta}^{[a]} + \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}^{-1}\mathbf{g}^{[a]}\right).
\end{aligned}$$

Therefore, the new fit for the parameter estimator can be expressed as

$$\boldsymbol{\theta}^{[a+1]} = \left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\mathbf{z}^{[a]}, \tag{A.2}$$

where $\mathbf{z}^{[a]} = \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\boldsymbol{\theta}^{[a]} + \boldsymbol{\epsilon}^{[a]}$ with $\boldsymbol{\epsilon}^{[a]} = \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}^{-1}\mathbf{g}^{[a]}$.

Appendix B. Gradient and Fisher information matrix

In this appendix, we derive the gradient vector and an approximation to the Fisher information matrix. The description to follow is also presented in Van den Hout (2017).

Given piecewise-constant intensities, the likelihood contribution for an observed time interval $(t_1, t_2]$ is defined using a constant generator matrix $\mathbf{Q} = \mathbf{Q}(t_1)$. For the eigenvalues of $\mathbf{Q}$ given by $\mathbf{b} = (b_1, ..., b_D)$, define $\mathbf{B} = \text{diag}(\mathbf{b})$. Given matrix $\mathbf{A}$ with the eigenvectors as columns, the eigenvalue decomposition is $\mathbf{Q} = \mathbf{ABA}^{-1}$. The transition probability matrix $\mathbf{P}(t) = \mathbf{P}(t_1, t_2)$ for elapsed time $t = t_2 - t_1$ is given by

$$\mathbf{P}(t) = \mathbf{A} \text{ diag}\left(e^{b_1 t}, ..., e^{b_D t}\right) \mathbf{A}^{-1}.$$

As described in Kalbfleisch and Lawless (1985), the derivative of $\mathbf{P}(t)$ can be obtained as

$$\frac{\partial}{\partial \theta_k}\mathbf{P}(t) = \mathbf{A}\mathbf{V}_k\mathbf{A}^{-1},$$

where $\mathbf{V}_k$ is the $D \times D$ matrix with (l, m) entry

$$\begin{cases} g_{lm}^{(k)} \left[\exp(b_l t) - \exp(b_m t)\right] / (b_l - b_m) & l \neq m \\ \\ g_{ll}^{(k)} t \exp(b_l t) & l = m, \end{cases}$$

where $g_{lm}^{(k)}$ is the (l, m) entry in $\mathbf{G}^{(k)} = \mathbf{A}\partial\mathbf{Q}/\partial\theta_k\mathbf{A}^{-1}$.

Let $\mathbf{g}(\boldsymbol{\theta})$ denote the $q \times 1$ gradient vector. The kth entry of $\mathbf{g}(\boldsymbol{\theta})$ is given by

$$\sum_{i=1}^{N}\sum_{j=2}^{n_i} \frac{\partial}{\partial \theta_k} \log P(Y_{ij} = y_{ij} | Y_{ij-1} = y_{ij-1}) \,.$$

The Fisher information matrix is given by $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = I\!E\left[\mathbf{g}(\boldsymbol{\theta})\mathbf{g}(\boldsymbol{\theta})^\top\right]$, which can be estimated by defining the $q \times q$ matrix $\mathbf{M}(\boldsymbol{\theta})$ with (k, l) entry

$$\sum_{i=1}^{N}\sum_{j=2}^{n_i} \frac{\partial}{\partial \theta_k} \log P(Y_{ij} = y_{ij} | Y_{ij-1} = y_{ij-1}) \frac{\partial}{\partial \theta_l} \log P(Y_{ij} = y_{ij} | Y_{ij-1} = y_{ij-1}) \,.$$

References

Cook, R.J., Lawless, J.F., 2018. Multistate Models for the Analysis of Life History Data. London: Chapman & Hall/CRC.

Cox, D.R., Miller, H.D., 1965. The Theory of Stochastic Processes. London: Chapman & Hall.

Eilers, P.H., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Statistical Science 11, 89–102.

Van den Hout, A., 2017. Multi-state survival models for interval-censored data. Boca Raton: CRC/Chapman & Hall.

Van den Hout, A., Matthews, F.E., 2008. Multi-state analysis of cognitive ability data: A piecewise-constant model and a Weibull model. Statistics in Medicine 27, 5440–5455.

Jackson, C.H., 2011. Multi-state models for panel data: The msm package for R. Journal of Statistical Software 38, 1–29.

Jackson, C.H., 2016. Flexsurv: A platform for parametric survival modelling in R. Journal of Statistical Software 70, 1–33.

Jackson, C.H., Sharples, L.D., Thompson, S.G., Duffy, S.W., Couto, E., 2003. Multistate markov models for disease progression with classification error. Journal of the Royal Statistical Society: Series D (The Statistician) 52, 193–209.

Joly, P., Commenges, D., 1999. A penalized likelihood approach for a progressive three-state model with censored and truncated Data: Application to AIDS. Biometrics 55, 887–890.

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, 433–443.

Kalbfleisch, J., Lawless, J.F., 1985. The analysis of panel data under a Markov assumption. Journal of the American Statistical Association 80, 863–871.

Machado, R.J.M., 2018. Penalised Maximum Likelihood Estimation for Multi-State Models. PhD thesis, University College London.

Machado, R.J.M., Van den Hout, A., 2018. Flexible multi-state models for interval-censored data: Specification, estimation, and an application to ageing research. Statistics in Medicine , 1–13.

Marioni, R.E., Valenzuela, M.J., Van den Hout, A., Brayne, C., Matthews, F.E., et al., 2012. Active cognitive lifestyle is associated with positive cognitive health transitions and compression of morbidity from age sixty-five. PLoS One 7, e50940.

Marra, G., Radice, R., Bärnighausen, T., Wood, S.N., McGovern, M.E., 2017. A Simultaneous Equation Approach to Estimating HIV Prevalence With Nonignorable Missing Responses. Journal of the American Statistical Association 112, 484–496.

Radice, R., Marra, G., Wojtyś, M., 2016. Copula regression spline models for binary outcomes. Statistics and Computing 26, 981–995.

Robitaille, A., van den Hout, A., Machado, R.J., Bennett, D.A., Čukić, I., Deary, I.J., Hofer, S.M., Hoogendijk, E.O., Huisman, M., Johansson, B., et al., 2018. Transitions across cognitive states and death among older adults in relation to education: A multistate survival model using data from six longitudinal studies. Alzheimer's & Dementia .

Titman, A.C., 2011. Flexible nonhomogeneous Markov models for panel observed data. Biometrics 67, 780–787.

Titman, A.C., Sharples, L.D., 2010. Model diagnostics for multi-state models. Statistical Methods in Medical Research 19, 621–651.

Wood, S., 2006. Generalized additive models: An introduction with R. CRC press.

Wood, S.N., 2007. The mgcv package. www. r-project. org .

Wu, X., Sickles, R., 2018. Semiparametric estimation under shape constraints. Econometrics and Statistics 6, 74–89.