Penalised maximum likelihood estimation for multi-state models

Robson José Mariano Machado

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Statistical Science

University College London

October 17, 2018

I, Robson José Mariano Machado, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Multi-state models can be used to analyse processes where change of status over time is of interest. In medical research, processes are commonly defined by a set of living states and a dead state. Transition times between living states are often interval censored. In this case, models are usually formulated in a Markov processes framework. The likelihood function is then constructed using transition probabilities. Models are specified using proportional hazards for the effect of covariates on transition intensities. Time-dependency is usually defined by parametric models, which can represent a strong model assumption. Semiparametric hazards specification with splines is a more flexible method for modelling time-dependency in multi-state models. Penalised maximum likelihood is used to estimate these models. Selecting the optimal amount of smoothing is challenging as the problem involves multiple penalties. This thesis aims to develop methods to estimate multi-state models with splines for interval-censored data. We propose a penalised likelihood method to estimate multi-state models that allow for parametric and semiparametric hazards specifications. The estimation is based on a scoring algorithm, and a grid search method to estimate the smoothing parameters. This method is shown using an application to ageing research. Furthermore, we extend the proposed method by developing a computationally more efficient method to estimate multi-state models with splines. For this extension, the estimation is based on a scoring algorithm, and an automatic smoothing parameters selection. The extended method is illustrated with two data analyses and a simulation study.

Acknowledgements

I would like to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil for funding this research. I would like to express my sincere gratitude to Dr Ardo van den Hout for his thorough supervision of my PhD. In particular, I am thankful for all the opportunities he provided, for motivating me and for his patience. I would like to thank Dr Giampiero Marra for providing good directions to this research. I am thankful to Dr Aidan O'Keeffe and Dr Andrew Titman for their valuable comments that significantly improved this thesis. I thank Dr Boo Johansson for permission to use the OCTO data, a connection that was made possible by Dr Graciela Muniz-Terrera to whom I am also thankful. In addition, I wish to thank the team of researchers behind the ELSA study. I would like to thank my friends and colleagues from University College London for the time spent together and exchange of ideas.

Finally, I am thankful to my family and friends for their support in general and throughout my academic path, especially to my parents for making me independent. Special thanks to Verena for making everything easier during the development of this research, for providing support in difficult moments and for celebrating the good ones.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research aim and overview

Multi-state models are a broadly applicable approach to analysing longitudinal data, where change of status over time is of interest. In medical research, these statuses are usually defined by the severity of a disease or condition of subjects in a follow-up study. Some standard examples include, dementia research in which interest lies in the decline of cognitive function, post-transplantation chronic disease studies, which aim to investigate disease progression after transplant, and human immunodeficiency virus (HIV), where modelling the transition of subjects across disease stages is of interest. Data deriving from these medical conditions share similar characteristics inherent to how data are obtained. Specifically, time of change in a medical condition cannot always be observed exactly, instead changes are recorded at pre-specified follow-up times. In this case, transition times are said to be interval-censored. On the other hand, if a dead state is defined, time of death is usually known exactly. Multi-state models for data with these characteristics are the focus of this thesis.

Multi-state models for interval-censored data are commonly formulated in a Markov processes framework (Kalbfleisch and Lawless, 1985). The Markov property states that the future of the process only depends on the current state. Models can be specified through the transition intensities (also called hazards), which represent the instantaneous risks of moving across states. To facilitate estimation, a

time homogeneous Markov process is usually assumed (Kalbfleisch and Lawless, 1985; Jackson, 2011). In this case, the hazards are assumed to be constant over time. For a wide range of applications, 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 (Van den Hout, 2017). However, the functional forms underlying the hazards are often unknown and parametric models can be too restrictive.

Flexible multi-state models can be obtained with smooth nonparametric hazard functions specification. In this case, the hazards are specified in terms of splines function basis. Splines are polynomial functions, which allow for flexible modelling De Boor (1978). In particular, they enable modelling time-dependent hazards without making strong model assumptions. For each hazard function, an extra parameter is employed to control model smoothness. These parameters are commonly called smoothing parameters. In the frequentist context, penalised maximum likelihood is used to estimate the models. Estimation of multi-state models with splines can be carried out in two challenging steps. First, for fixed values of smoothing parameters, penalised maximum likelihood estimation is used to obtain estimates for the model parameters. Second, given the estimates of model parameters in the first step, we aim to define a stable and efficient method to estimate the optimal values for the smoothing parameters. These two steps are iterated until a convergence criterion is met. At the present time, methods available cannot fully address the problem of estimating flexible multi-state models in the presence of interval censoring.

This thesis aims to develop a new efficient method for estimating multi-state models with splines in the presence of censoring. The focus is on observation schemes in which the transition times between living states are interval-censored and times into the dead (or absorbing) state are known exactly (or right-censored).

In the remainder of this chapter, some special features of multi-state modelling are presented. This includes a description of basic concepts of survival analysis that are used throughout this thesis. In addition, the literature on techniques for

parametric and semiparametric multi-state models is reviewed. This also includes a review of smoothing methods useful for the development of methods in this thesis. Finally, we introduce data used to illustrate methods studied and developed in this thesis.

Chapter 2 follows this introduction by presenting the theory of multi-state modelling. Specifically, we discuss parametric model specification and estimation. A detailed description of a scoring algorithm to obtain the maximum likelihood estimates is discussed. The method is then illustrated through an application to data for post-heart transplantation patients, and data for decline of cognitive function.

Chapter 3 then extends parametric multi-state models by allowing for transition-specific hazards specification with splines. Firstly, we provide an introduction to smoothing methods that will be used for the remainder of the thesis. In particular, we present cubic regression splines and P-splines. We then present our method for specification and estimation of multi-state models with splines, that uses a scoring algorithm for estimating the model parameters, and a grid search for selecting the optimal amount of smoothing. This method is then illustrated with an application to ageing research.

Chapter 4 proposes an unifying and efficient framework for estimating multi-state models with splines in the presence of censoring. A simulation study is carried out to analyse the performance of the method presented. Subsequently, the method is illustrated with an application to data used in Chapter 2.

The final chapter provides a discussion of the methods and applications developed in this thesis, and indicates areas for future research.

1.2 Special features of multi-state processes

In this section, a number of different multi-state models are presented in order to illustrate some essential features of these models. In particular, we discuss some model structures and observational patterns commonly found in medical research. The discussion presented in this section is mainly based on the papers by Andersen and Keiding (2002) and Commenges (2002).

1.2.1 Model structure

Graphically, multi-state models can be illustrated using diagrams with boxes representing the states and with arrows between the states representing the possible transitions. A state is said an *absorbing* state if further transitions cannot occur from that state. A *transient* state is a state that is not absorbing, meaning that at least one transition is possible from that state.

Survival model

A simple multi-state model can be defined for survival data in which individuals are followed-up until the event death (or censoring) occurs. The two-state model is then defined by the transient State 1 (alive) and the absorbing State 2 (death). This model is illustrated in Figure 1.1.

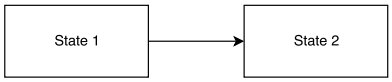

Figure 1.1: A two-state survival model

Unidirectional model

Unidirectional models extend the survival model by defining a set of transient states before an absorbing State D (Figure 1.2) (Titman, 2008). For this model, individuals start at a transient state and can move forward to the next state until the absorbing state. Cook and Lawless (2002) use these models for the analysis of repeated events data.

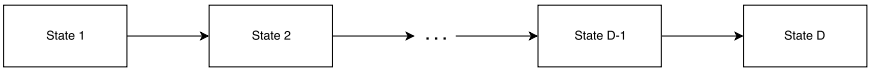

Figure 1.2: An unidirectional model

The illness-death model

The illness-death model is defined by a disease-free state (State 1), from which subjects can move into the dead state (State 3), or move into a diseased state (State 2). Once in State 2, subjects can move back to State 1 or move into State 3, see Figure 1.3.

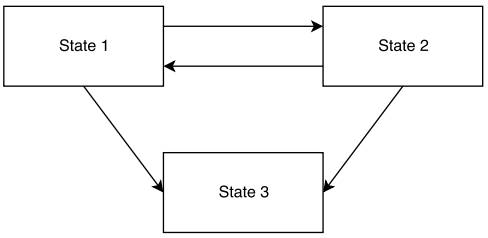

Figure 1.3: An illness-death model

Competing risks model

In the context of medical research, competing risks models are defined by one transient state (State 1: alive) and a number, k, of absorbing states, State h, for $h = 1, \ldots, k$ corresponding to "death from cause h". Figure 1.4 illustrates the model for $k = 2$.

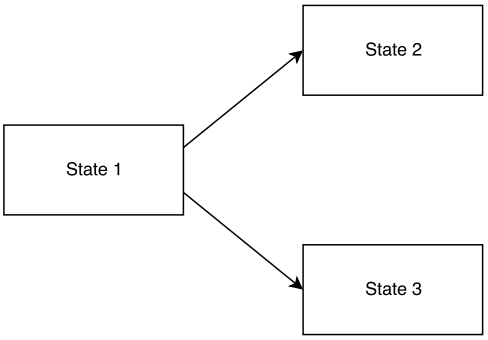

Figure 1.4: A competing risks model

1.2.2 Observational patterns

As discussed in Commenges (2002), the observations are often incomplete in the sense that it is not possible to observe the whole population of interest, and it is not possible to observe the process continuously over time. The first problem can be addressed by drawing a representative sample of the population. The second problem leads to what is called censored observations. Commenges (2002) provides a thorough discussion on how to approach different types of censoring in multi-state processes. We next describe right and interval-censoring as these are the focus of this thesis.

Consider first that a multi-state process is observed in continuous time from time t_0 until time $t_0 + c$. If the state of the process at time $t_0 + c$ is an absorbing

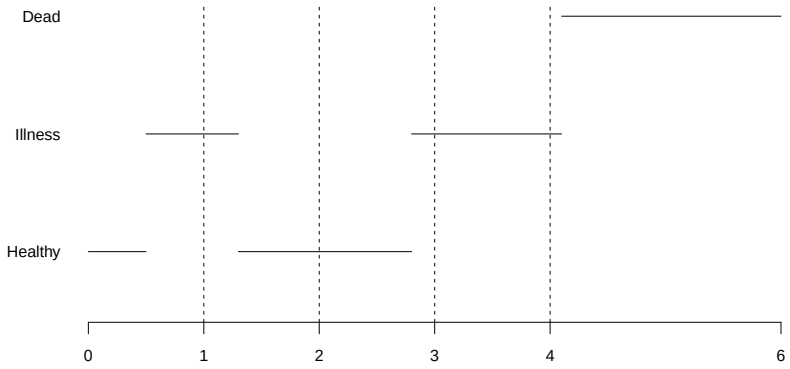

Figure 1.5: The trajectory of an illness-death process. Dashed vertical lines represent the follow-up times. Solid horizontal lines represent the state occupancy over time

state, then the process is observed completely. If not, then some transition times are said to be right-censored.

For many applications, a multi-state process is observed at only a finite number of times. In this case, the exact time of transitions are only known to have occurred within a time interval. In this case, the transition times are said to be interval censored. This type of censoring is common in cohort studies, where states are recorded at fixed visit times. If a dead state is defined, the time of death is usually known exactly. In this case, we have a mix of discrete and continuous time observations. This thesis focus on methods for these pattern of observations. Figure 1.5 illustrates the trajectory of an illness-death process over time. The vertical dashed lines represent the follow-up times at which state occupancy is recorded. The horizontal lines represent the state occupancy over time. The process moves back and forth between the living states, subsequently moves from the illness state into the dead state.

1.3 Literature review

For the analysis of time-homogeneous multi-state processes, Kalbfleisch and Lawless (1985) developed a general procedure for obtaining maximum likelihood estimates of the model parameters. The method is based on a scoring algorithm that uses an approximation to the second order derivatives of the log-likelihood function. Kay (1986) proposes a similar method that calculates the first and second or-

der derivatives of some particular multi-state processes. Kay also provides methods for hypothesis testing and model diagnostics. Satten and Longini (1996) developed a method for fitting these models when states are subject to measurement errors. Jackson (2011) presented the msm package in R for analysing time homogeneous multi-state processes for interval-censored data.

For time-dependent multi-state processes, Kalbfleisch and Lawless (1985) suggested using two methods. The first method uses piecewise-constant hazards to approach time-dependent processes. In this case, the hazards are constant within specified intervals, but can change for different intervals, see also Kay (1986). For many applications, it is not reasonable to assume that the underlying hazards are piecewise-constant functions. Also, the total number of parameters in the model can become large with the number of transition specific-hazards. The second method focuses on a special case in which the non-homogeneity is due to a time-varying multiplicative change in the matrix of transition intensities. In this case, it is possible to find a transformation function so that the resulting process is time homogeneous. Hubbard et al. (2008) investigated this method and proposed a flexible approach for estimating the transformation functions. Titman (2011) uses a numerical approximation to calculate the transition probabilities at the level of the corresponding differential equations. A disadvantage of this method is that estimation can become computationally expensive if many time-varying covariates are included. Jackson (2011) suggests using a piecewise-constant approximation for parametric models, which is employed to obtain the transition probabilities for the likelihood function. Van den Hout (2017) provides a general method for estimating multi-state models for interval-censored data. Focus is given to time-dependent parametric models, such as, Gompertz and Weibull distributions. Given a piecewise-constant approximation to the parametric hazards, a scoring algorithm is used for estimating the models. Further literature on time-dependent multi-state models includes Omar et al. (1995), Van den Hout and Matthews (2008a) and Van den Hout and Matthews (2008b).

Even though multi-state models with splines for known transition times are out

of the scope of this research, a literature review of these models can be important for the development of this work. Fahrmeir and Klinger (1998) approached these models in a counting process framework. Transition intensities are specified in terms of spline functions. Estimation is based on a backfitting scheme with internal smoothing parameter selection by AIC optimisation. The application is to human sleep data with a discrete set of sleep states. Measurements are made every 30 seconds for a group of 30 patients. A Bayesian approach to this model and data is given in Kneib and Hennerfeind (2008). In this case, the inferential procedures developed allow for simultaneous estimation of model and smoothing parameters. Sennhenn-Reulen and Kneib (2016) developed an estimation procedure based on a structured lasso penalisation for multi-state models. The aim of their research is to identify covariate effect coefficient equal to zero. Baseline transition intensities are specified with piecewise-constant models, or unspecified and equal across all transitions. The R package Flexsurv is mainly designed for modelling time-to-event data (Jackson, 2016). It can fit any parametric time-to-event distribution and the spline models of (Royston and Parmar, 2002). This package can also be used for fitting some multi-state model by writing data in a survival format. Models can be specified with a rage of parametric and nonparametric shapes.

In the context of multi-state models with splines for interval-censored data, a penalised approach for an unidirectional three-state model is used in Joly and Commenges (1999). Estimation is performed with an algorithm which uses derivatives of the penalised log-likelihood. The smoothing parameters are selected using a grid search with cross-validation. Joly et al. (2002) use the same approach for an illness-death without recovery model. Joly et al. (2009) further extend their method for a five-state model without recovery. These methods require 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 backward transitions. In addition, their methods can be computationally intractable for models with multiple smoothing parameters, as the grid search method requires models to be fitted for all possible combination of smoothing parameters values given in the grid.

The method proposed in Titman (2011) allows for nonparametric hazard specifications with B-splines; however, the log-likelihood is maximised without penalisation. If a large set of B-splines basis is used to model a transition-specific hazard, models can become unidentifiable.

Efficient smoothing parameter selection methods are important for practical multi-state modelling with splines. We next discuss some relevant literature for automatic smoothing parameter estimation that will be useful for defining an automatic and efficient algorithm to estimate multi-state models with splines. Gu and Wahba (1991) developed an efficient Newton method for multiple smoothing parameter selection. Wood (2000) extends their method for multiple smoothing parameter selection in generalized ridge regression problems. Wood (2004) provides a stable and efficient method that improves the method developed in Wood (2000). These methods are widely used in generalised additive models. However, they can be further extended for more general settings. Radice et al. (2016) proposes a method for multiple smoothing parameter estimation for copula regression. Their method is general, but lead to numerical instability for some applications. Marra et al. (2017) developed a more general and stable method that can be estimate multiple smoothing parameters using only the gradient and Hessian (or Fisher information matrix). This method is general and can be used in a variety of settings.

1.4 Survival analysis

Some concepts of time-to-event analysis can be used and extended to multi-state modelling. In this section, we describe important features of these concepts that will be useful for the development of this thesis. Time-to-event analysis aims to describe the analysis of data in the form of a well-defined time origin until the occurrence of some particular event. In medical research, the time origin will often correspond to beginning of a treatment and the event of interest can be relief of pain, the recurrence of symptoms, or death. If the event is death of a patient, then time-to-event data are commonly called survival data. In this thesis, we focus on examples that include a dead state, and we shall refer to time-to-event data as survival data.

This section is partly based on Collett (2015).

Let T be the random variable associated with the survival time of patients, which can take any non-negative values. The distribution function of T is given by

$$F(t) = P(T < t) = \int_0^t f(u)du, \tag{1.1}$$

where $f(t)$ represents the *probability density function* of T. Equation (1.1) represents the probability that survival is less than some value t.

The *survival function*, $S(t)$, is defined to be the probability that the survival time is greater than or equal to t,

$$S(t) = P(T \geq t) = 1 - F(t), \tag{1.2}$$

where $F(t)$ is as defined in (1.1).

The *hazard function* is widely used to express the risk or hazard of an event occurring at some time t. This function is obtained from the probability that an individual dies at time t, conditional on them having survived to that time. Formally, the hazard function, $h(t)$, is given by

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T \leq t + \delta t | T \geq t)}{\delta t} \right\}. \tag{1.3}$$

Equations (1.1), (1.2) and (1.3) provide that the hazard function can be expressed as

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.4}$$

The *cumulative hazard function* is used to express the cumulative risk of an event occuring by time t. Formally, the cumulative hazard function, $H(t)$, is given by

$$H(t) = \int_0^t h(u)du. \tag{1.5}$$

From Equation (1.4), it follows that $h(t) = -\frac{d}{dt}\{\log S(t)\}$, which implies that $S(t) = \exp\{-H(t)\}$. Therefore, the survival function can also be expressed in terms of the cumulative hazard function.

Empirical non-parametric methods are useful to summarise the survival times for individuals in a particular group. Suppose that for a single sample of survival times none of the observations is censored. The *empirical survival function* is given by

$$\widehat{S}(t) = \frac{\text{Number of individuals with survival times} \geq t}{\text{Number of individuals in the data}}. \tag{1.6}$$

The empirical survival function is equal to unity for values of t before the first death time, and zero after the final death time. The estimated survivor function, $\widehat{S}(t)$, is assumed to be constant between two adjacent death times, and so a plot of $\widehat{S}(t)$ against t is a step-function. The function decreases immediately after each observed survival time.

Suppose that there are n individuals with survival times at $t_1, \ldots, t_n$. These observations can be right-censored and more than one death can occur at a given time. Suppose that there are r deaths among the individuals, where $r \leq n$. Let $t_{(1)}, \ldots, t_{(r)}$ be the r ordered death times, n_j the number of individuals who are alive just before $t_{(j)}$, including those who are about to die, and d_j the number of individuals who die at $t_{(j)}$. The Kaplan-Meier estimate of the survival function is given by

$$\widehat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \tag{1.7}$$

for $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, \ldots, r$, with $\widehat{S}(t) = 1$ for $t < t_{(1)}$, and where $t_{(r+1)}$ is taken to be ∞. If the largest observation is a censored survival time, t^*, then $\widehat{S}(t)$ is undefined for $t > t^*$. If the largest observed survival time, $t_{(r)}$, is an uncensored observation, $n_r = d_r$, and so $\widehat{S}(t) = 0$ for $t \geq t_{(r)}$. A plot of the Kaplan-Meier estimate of the survival function is a step-function, in which the estimated survival probabilities are constant between adjacent death times and decrease at each death time. This thesis uses the Kaplan-Meier estimate of the survival function for model validation.

1.4.1 The exponential distribution

The probability density function of a random variable T that has an exponential distribution with parameter $\lambda > 0$ is given by

$$f(t) = \lambda \exp^{-\lambda t}, \tag{1.8}$$

for $0 \leq t < \infty$. The survival function is then given by

$$\begin{aligned} S(t) &= 1 - \int_0^t f(t)dt \\ S(t) &= \exp(-\lambda t). \end{aligned} \tag{1.9}$$

Using the relation $h(t) = -\frac{d}{dt}\log(S(t))$, the hazard function is given by

$$h(t) = \lambda, \tag{1.10}$$

for $t \geq 0$. Therefore, if the hazard is constant over time. For many applications, it may be more reasonable to consider time-dependent hazard functions.

1.4.2 The Gompertz distribution

The probability density function of a random variable T that has a Gompertz distribution with parameters $\lambda > 0$ and $\theta \in \mathbb{R}$ is given by

$$f(t) = \lambda e^{\theta t} \exp\left\{\frac{\lambda}{\theta}(1 - e^{\theta t})\right\}, \tag{1.11}$$

for $0 \leq t < \infty$. The survival function of the Gompertz distribution is given by

$$S(t) = \exp\left\{\frac{\lambda}{\theta}(1 - e^{\theta t})\right\}. \tag{1.12}$$

The hazard function of the Gompertz is given by

$$h(t) = \lambda e^{\theta t}, \tag{1.13}$$

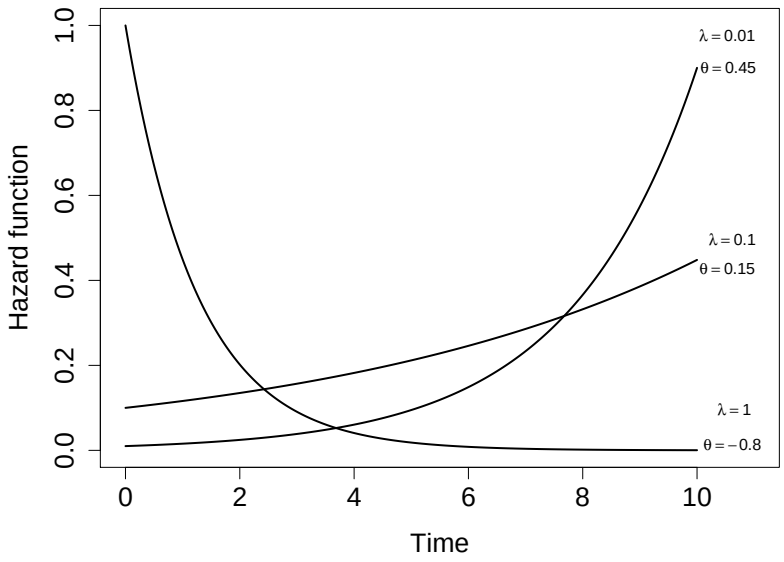

Figure 1.6: Gompertz hazard functions for scale and shape parameters $(\lambda, \theta) = (1, -0.8), (0.1, 0.15)$ and $(0.01, 0.45)$

for $t \geq 0$. The parameters θ and λ are the shape and scale parameters, respectively. For the case in which $\theta = 0$, the hazard function is constant and equal to λ, and the survival times then have an exponential distribution. Positive values of θ lead to increasing hazards, while negative values of θ lead to decreasing hazards. As an illustration, the hazard function for Gompertz distributions with parameters $(\lambda, \theta) = (1, -0.8), (0.1, 0.15)$ and $(0.01, 0.45)$ are shown in Figure 1.6.

1.4.3 The Weibull distribution

The probability density function of a random variable T that has a Weibull distribution, with parameters $\lambda > 0$ and $\gamma > 0$ is given by

$$f(t) = \lambda \gamma^{\gamma - 1} \exp(-\lambda t^{\gamma}), \tag{1.14}$$

for $0 \leq t < \infty$. The survival function of the Weibull distribution is given by

$$S(t) \quad = \quad \exp(-\lambda t^{\gamma}). \tag{1.15}$$

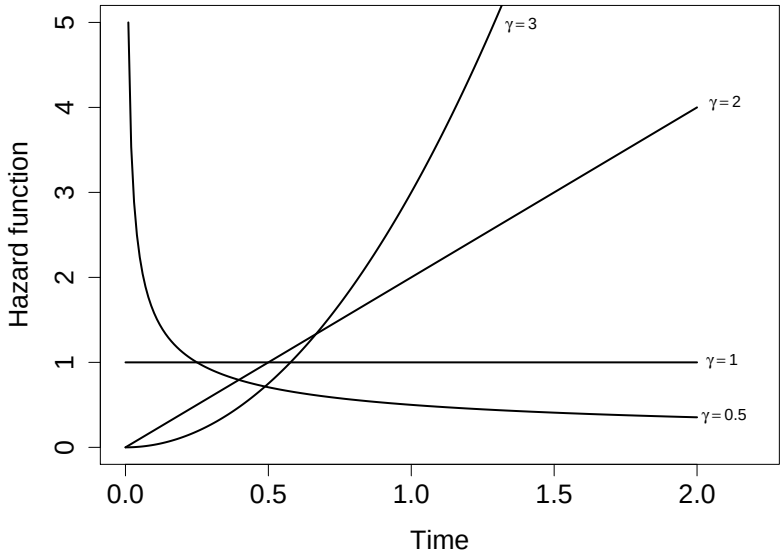

Figure 1.7: Weibull hazard function for scale parameter $\lambda = 1$ and the shape parameters $\gamma = 0.5,\ 1,\ 2$ and 3

The corresponding hazard function is given by

$$h(t) = \lambda \gamma t^{\gamma - 1} \tag{1.16}$$

for $t \geq 0$. For the case in which $\gamma = 1$, the hazard function is constant over time, and the survival times have an exponential distribution. The shape of the hazard function is controlled by the parameter γ, which is then defined as the *shape parameter*, while λ is a *scale parameter*. A plot of the Weibull hazard function for distributions with $\lambda = 1$ and $\gamma = 0.5, 1, 2$ and 3 is shown in Figure 1.7.

1.4.4 The lognormal distribution

A random variable T has a lognormal distribution, with parameters μ and σ, if the random variable $\log T$ has a normal distribution with mean μ and variance σ^2. The probability density function of T is given by

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} t^{-1} \exp\left\{-(\log t - \mu)^2 / 2\sigma^2\right\}, \tag{1.17}$$

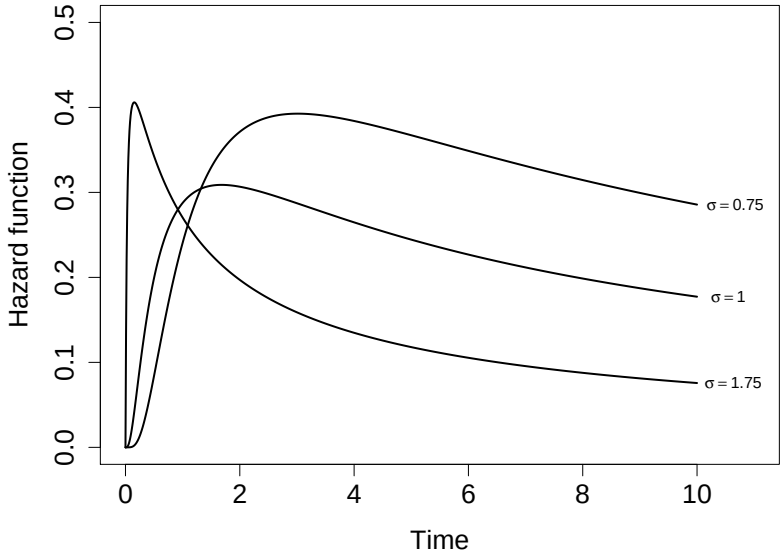

Figure 1.8: Lognormal hazard functions for $\mu = 1$ and $\sigma^2 = 0.75,\ 1$, and 1.75

for $0 \leq t < \infty$ and σ. The survival function of the lognormal distribution is

$$S(t) \quad = \quad 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \tag{1.18}$$

where $\Phi(\cdot)$ is the standard normal distribution function, given by

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-u^2/2) du. \tag{1.19}$$

The hazard function can be found from the relation $h(t) = f(t)/S(t)$. This function is zero at $t = 0$, increases to a maximum and then decreases to zero as t tends to infinity. The hazard and survival functions are given in terms of integrals limits, which makes the use of this model difficult for applications. However, the flexibility of this model is attractive and they will be useful in Chapter 4 to simulate transition times from a distribution function that leads to a non-linear shape of hazard functions. For illustration, the hazard function for lognormal distributions with parameters $\mu = 1$ and $\sigma^2 = 0.75, 1$, and 1.75 are shown in Figure 1.6.

1.5 Simulation of multi-state processes

In this section, we describe how to simulate multi-state processes using the inversion method (Bender et al., 2005). For a fixed transition from state r to state s, $r \neq s$, the random variable T represents the time to event . If the cumulative distribution function for leaving state r to state s is given by $F(t) = 1 - S(t)$, then $U = F(T)$ has a uniform distribution on the interval $[0,1]$, denoted by $U \sim U[0,1]$. Moreover, if a random variable has distribution $U \sim U[0,1]$ then $1-U \sim U[0,1]$ as well. It means that $S(T) \sim U[0,1]$. Therefore, if the function $H(t)$ can be inverted, the time to event from state r to state s can expressed as

$$T = H^{-1}[-\log(U)\exp(\beta_{rs}^{\top}\mathbf{z})] \tag{1.20}$$

where $U \sim U[0,1]$.

Suppose there are $m-1$ competing intensities for leaving state r. The time to next event can be obtained by taking $T = \min(T_1, \ldots, T_{m-1})$, where T_i is obtained applying (1.20).

For the calculation of subsequent event times, we use that the survival function conditional on entering the state r at time $t_0 > 0$ is given by

$$\begin{aligned} S(t|t_0) &= P(T > t|T > t_0) \\ &= \frac{S(t)}{S(t_0)}. \end{aligned} \tag{1.21}$$

1.6 Examples

This section further discusses special features of multi-state processes through examples. Data from these examples will then be used to illustrate the statistical methods presented in subsequent chapters.

1.6.1 Cardiac allograft vasculopathy data

Cardiac allograft vasculopathy (CAV) is a narrowing of the arterial walls and one of the main cause of death in heart transplantation patients. The data are a series of approximately yearly angiographic examinations of heart transplant recipients.

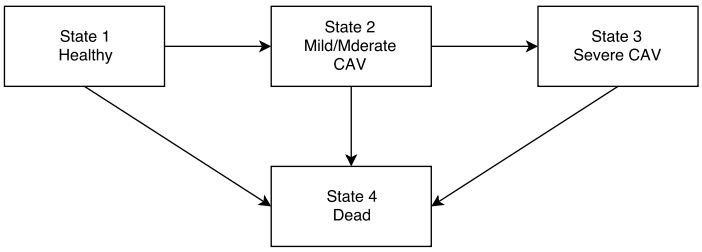

Figure 1.9: Four-state model for disease progression after transplant for the CAV data

The data come from Papworth Hospital U.K. The data contain 3217 rows which are grouped by 622 patients and ordered by years after transplant. For the CAV data, baseline means years since the beginning of the study.

Of interest is the onset of CAV after transplantation. The state at each follow-up time is a grade of CAV which can be normal, mild, moderate or severe. Dead is the absorbing state and time of death is known within one day. Three living states are defined by CAV severity: State 1, 2, and 3, for Normal, Mild/Moderate, and Severe, respectively. An additional State 4 is defined as the Dead state, see Figure 1.9.

The interval-censored multi-state process is summarised by the frequencies in Table 1.1. The code 99 represents right-censored observations. As it can be seen, some backward transitions are recorded. However, the narrowing process is assumed to be biologically irreversible. There are essentially two ways to approach this problem. First, it is possible to fit multi-state models with misclassification of states (Jackson et al., 2003). Such an approach is out of the scope of the current research. Second, the data can be redefined for the history of CAV. In this case, the states are classified as Healthy (1) if the patient has not developed the disease, Mild/Moderate (2) if the patient has developed mild/moderate or Severe (3) if the patient has developed severe CAV and Dead (4) if the patient has died. Then, the data are consistent with the model in Figure 1.9.

The CAV data include several covariates which may be used to analyse patient's progression across states. These include recipient age at examination, donor age, the sexes of recipient and donor, and primary diagnosis of ischaemic heart disease (IHD). Sharples et al. (2003) showed in an analyse of these data that IHD and

Table 1.1: State table for the CAV data: number of times each pair of states was observed at successive observation times. The three living states are defined by CAV severity

	To				
From	1	2	3	4	99
1	1367	204	44	148	276
2	46	134	54	48	69
3	4	13	107	55	26

donor age are major risk factors of disease onset.

1.6.2 English longitudinal study of ageing data

The English Longitudinal Study of Ageing (ELSA) baseline (1998-2001) is a representative sample of the English population aged 50 and older. ELSA contains information on health, economic position, and quality of life. Data from ELSA can be obtained via the Economic and Social Data Service (`www.esds.ac.uk`). There are 11932 individuals in the ELSA baseline. The following description of these data can also be found in Van den Hout (2017).

Of interest is the change of cognitive function in older population. For the analysis in this thesis, a random sample of size $N = 1000$ is taken from ELSA. Of these 1000 individuals, 205 died during the follow-up with age at death available. Because ELSA data are publicly available, measures have been taken by the data provider to prevent identification of the individuals. One of those measures is the censoring of ages above 90 years. In the sampling of the subset of $N = 1000$, individuals who were 90 years or older at baseline were ignored, where baseline means entry in the study. The sample has 544 women and 456 men. Highest educational qualification is dichotomised according to years of formal education: fewer than ten versus ten or more. There are 558 individuals with fewer than ten years of education.

The focus is on the number of words remembered in a delayed recall from a list of ten. The score on this test is equal to the number of words remembered, i.e., score $\in \{0, 1, 2, \cdots, 10\}$. It is of interest to explore the effect of age and gender on cognitive change over time when controlling for education. Four living states are

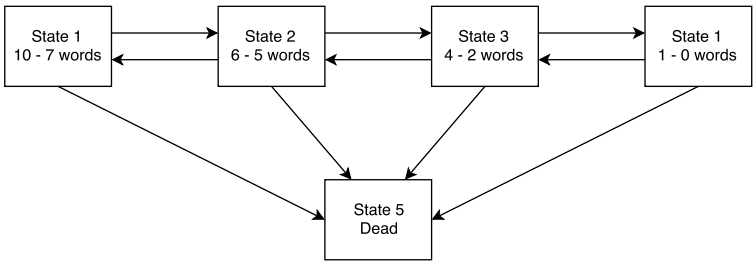

Figure 1.10: Five-state model for longitudinal data in ELSA on number of words remembered in a recall

defined by the number of words an individual can remember: State 1, 2, 3, and 4, for the number of words $\{7,8,9,10\}$, $\{6,5\}$ $\{4,3,2\}$, and $\{1,0\}$, respectively. An additional State 5 is defined as the dead state, see Figure 1.10.

1.6.3 Origins of variance in the oldest-old data

The origins of variance in the oldest-old (OCTO) study included dizygotic (DZ) and monozygotic (MZ) twin pairs aged 80 years of age and older. The sample was selected from older adults in the population-based Swedish Twin Registry. Older adults participating in the study were tested in their residence by nurses. Informed consent was obtained from each participant. Five cycles of longitudinal data were collected at two year intervals. The initial sample consisted of 702 individuals (351 same-sex pairs). The final analysis included 694 participants.

Of interest is the effects of age on cognitive function. The mini-mental state examination (MMSE) is used to assess global cognitive functioning of participants at each time point (Folstein et al., 1975). The state at each follow-up time is a classification of respondents in terms of MMSE, which can be normal MMSE ($27 \leq$ MMSE ≤ 30), mild MMSE impairment ($23 \leq$ MMSE ≤ 26) and severe MMSE impairment (MMSE ≤ 22). The MMSE is not used to determine clinical diagnosis but as suggestive of mild cognitive impairment and dementia. These MMSE severity are used to define three living states: State 1, 2, and 3, for No cognitive impairment, Mild cognitive impairment, and Severe cognitive impairment, respectively. An additional State 3 is defined as the Dead state, see Figure 1.11.

The interval-censored multi-state process is summarised by the frequencies in Table 1.2. Even though there are some backwards transitions from State 3, severe

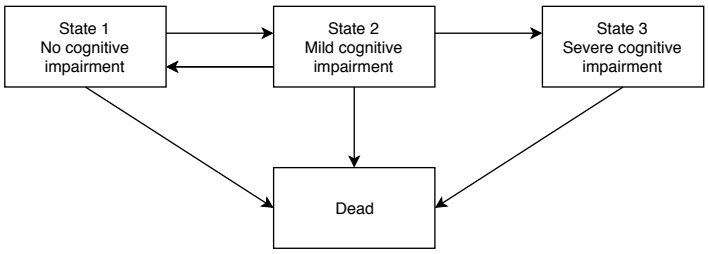

Figure 1.11: Four-state model for longitudinal data in OCTO

cognitive impairment is assumed to be irreversible. Here, the OCTO data are redefined for the history of severe cognitive impairment. In this case, if a participant has moved into State 3, they are only allowed to stay in State 3 or move into the dead state, see Figure 1.11.

The data include several covariates which may be used to analyse patient's decline of cognitive function and death. These include education, sex, and socioeconomic status (SES). Robitaille et al. (2018) investigate the effect of age and these covariates on the transitions between cognitive states and life.

1.7 Discussion

To conclude, estimation of time-dependent multi-state models for interval-censoring can be challenging. Most parametric models are specified with restrictive functional forms, such as, Gompertz and Weibull. Flexible parametric distributions can also be used, but obtaining the derivatives of the log-likelihood function for those models can be intractable. For the examples in Section 1.6, there is no clear information on good parametric shapes. Nonparametric hazards specification with splines provides a flexible approach for modelling time-dependent multi-state process. The methods developed so far focus on particular cases and are not feasible for many applications.

Table 1.2: State table for the OCTO data: number of times each pair of states was observed at successive observation times. The three living states are defined by grades of cognitive function

	To			
From	1	2	3	4
1	715	133	82	233
2	67	106	116	110
3	8	21	274	319

Chapter 2

Parametric multi-state models

This chapter introduces parametric multi-state models for interval-censored data. The general formulation builds up within a Markov processes framework, and models are specified through hazard functions. Maximum likelihood is used for estimation. Model selection and model validation are briefly discussed as they will be useful for comparing and validating models throughout this thesis. The method is illustrated with the CAV and OCTO data sets. These applications show the versatility of multi-state modelling for longitudinal data, but also highlight some limitations of parametric model specifications. The aim of this chapter is to present the general theory of multi-state models. All methods presented are standard and can be found in textbooks (Cox, 2017; Van den Hout, 2017). The applications are original, as the CAV and OCTO data sets have not been analysed with the formatting proposed in this chapter.

2.1 Continuous-time Markov processes

A *stochastic process* is a collection of random variables $\{Y(t)|t \in U\}$, where U is an index set that can take discrete or continuous values. The *state space*, $\mathscr{S}$, is the set of possible values of $Y(t)$, which can also be discrete or continuous. The focus here is on the case in which U is continuous and $\mathscr{S}$ is discrete, where the former is a set of model states and $Y(t)$ denotes the state of $\mathscr{S}$ occupied by the system at time t. The content presented in this section follows the presentation in Cox (2017).

Given the time points $t_1, \ldots t_n$, it is of interest to examine the joint distribu-

tion of $Y_1,\ldots,Y_n$, where $Y_j = Y(t_j)$ for $j = 1,\ldots,n$. Commonly, $\{Y(t)|t \in U\}$ is assumed to be a Markov process, which means that the future state of the process only depends on the current state. Formally, a continuous-time Markov process on the discrete states $\mathscr{S}$ is defined through a set of probabilities, $p_{rs}(t)$, such that,

$$p_{rs}(t,u) = P\big(Y(u+t) = s|Y(u) = r\big), \tag{2.1}$$

for $r,s \in \mathscr{S}$, $u \geq 0$ and $t \geq 0$. The Markov process $\{Y(t)|t \in U\}$ is time-homogeneous if the probability (2.1) only depends on the initial state, that is,

$$p_{rs}(t) = P\big(Y(t) = s|Y(0) = r\big). \tag{2.2}$$

The process is for now assumed to be time-homogeneous. The probabilities in (2.2) must satisfy

$$0 \leq p_{rs}(t) \leq 1, \tag{2.3}$$

$$p_{rk}(t) = \sum_s p_{rs}(u)p_{sk}(t-u) \quad (t > u), \tag{2.4}$$

$$\sum_s p_{rs}(t) = 1. \tag{2.5}$$

The matrix $\mathbf{P}(t)$ which contains these probabilities is called the *transition probability* matrix. Equation (2.4) is the *Chapman-Kolmogorov* equation for a time-homogeneous Markov process and can be written in matrix form as $\mathbf{P}(t) = \mathbf{P}(u)\mathbf{P}(t-u)$, with $\mathbf{P}(0) = \mathbf{I}$. This is useful for computing transition probabilities over a long-term time interval.

In applications, models are specified through the transition rates over a small time interval. The *transition intensities* (or *hazards*) from state r to state s are given by

$$q_{rs} = \lim_{\Delta t \to 0} \frac{P\big(Y(t+\Delta t) = s|Y(t) = r\big)}{\Delta t}, \tag{2.6}$$

for $r \neq s$. Notice that the $q'_{rs}s$ are constants. The matrix $\mathbf{Q}$ with off-diagonal entries q_{rs} and diagonal entries $q_{rr} = -\sum_{s\neq r} q_{rs}$ is called the *generator matrix*. If $q_{rr} = 0$,

the state r is called *absorbing*. As described in Section 1.6, data analysed in this thesis will include an absorbing state, which represents the dead state.

As an example, the generator matrix for the ELSA data described in Section 1.6.2 is given by

$$\begin{pmatrix} -(q_{12}+q_{15}) & q_{12} & 0 & 0 & q_{15} \\ q_{21} & -(q_{21}+q_{23}+q_{25}) & q_{23} & 0 & q_{25} \\ 0 & q_{32} & -(q_{32}+q_{34}+q_{35}) & q_{34} & q_{35} \\ 0 & 0 & q_{43} & -(q_{43}+q_{45}) & q_{45} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The rate at which individuals move, e.g., from state 1 into state 3 is zero, $q_{13} = 0$, because this transition is not allowed by the process. In order to move into state 3 from state 1 subjects must go first through state 2. The dead state 5 is the only absorbing state for this application.

For a given generator matrix, $\mathbf{Q}$, a Markov process is uniquely defined (Cox, 2017). The link between a generator matrix and its probability matrix is established by the forward and backward equations, which are given by

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}, \tag{2.7}$$

$$\mathbf{P}'(t) = \mathbf{Q}\mathbf{P}(t), \tag{2.8}$$

respectively. Given the initial condition $\mathbf{P}(0) = \mathbf{I}$, the solution to the differential equations in (2.7) and (2.8) is

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) \tag{2.9}$$

$$= \sum_{k=0}^{\infty} \mathbf{Q}^k \frac{t^k}{k!}, \tag{2.10}$$

where $\mathbf{Q}^0 = \mathbf{I}$. Because $\mathbf{Q}$ is finite, the series (2.10) is convergent and (2.9) is the unique solution of both backward and forward equations. Moler and Van Loan (2003) discuss a range of methods to calculate the exponential of a matrix. Here, the

transition probability matrix $\mathbf{P}(t)$ are computed using the eigen-decomposition of $\mathbf{Q}$. Let $b_1, \ldots, b_k$ represent the eigenvalues of $\mathbf{Q}$ and $\mathbf{A}$ the matrix with the eigenvectors as columns. For distinct eigenvalues, the matrix A is invertible, and the eigenvalue decomposition is $\mathbf{Q} = \mathbf{A}\text{diag}(b_1, \ldots, b_k)\mathbf{A}^{-1}$. The transition probability matrix $\mathbf{P}(t)$ for elapsed time t is given by

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

In this thesis, *multi-state processes* are Markov processes with transition probabilities depending on the current state and possibly on covariates.

For many applications, 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. The generator matrix is then a function of time, which means that the matrix $\mathbf{Q}(t)$ can vary over time. We next describe how to specify and approach time-dependent multi-state models.

2.2 Model representation

The hazard functions as defined in (2.6) can be generalised to the situation where the hazards depend on time and on the values of p explanatory variables, $x_1, \ldots, x_p$. The set of values of the explanatory variable are denoted by the vector $\mathbf{x} = (x_1, \ldots, x_p)^\top$. Let $q_{rs.0}(t)$ represent the hazard function for an individual for whom the values of all the explanatory variables that make up the vector $\mathbf{x}$ is zero. This function is called the *baseline hazard function*. A time-dependent hazard regression model can be written in the form

$$q_{rs}(t) = q_{rs.0}(t)\exp\left(\boldsymbol{\beta}_{rs}^\top \mathbf{x}\right), \tag{2.11}$$

where $\boldsymbol{\beta}_{rs} = (\beta_{rs.1}, \ldots, \beta_{rs.p})^\top$ is the vector of coefficients of the p explanatory variables acting on transition from state r to state s. We focus on the case where the explanatory variables are recorded at the time origin of the study. It is straightforward to extend the model to the situation where the values of the explanatory

variables change over time (Van den Hout, 2017).

Equation (2.11) shows that transition-specific time dependency can be introduced via baseline hazards. As presented in Section 1.4, examples of parametric functional forms for $q_{rs.0}(t)$ include

$$\text{exponential:} \quad q_{rs.0}(t) = \alpha_{rs} \qquad \alpha_{rs} > 0 \tag{2.12}$$

$$\text{Weibull:} \quad q_{rs.0}(t) = \alpha_{rs}\tau_{rs}t^{\tau_{rs}-1} \qquad \alpha_{rs}, \tau_{rs} > 0 \tag{2.13}$$

$$\text{Gompertz:} \quad q_{rs.0}(t) = \alpha_{rs}\exp(\xi_{rs}t) \qquad \alpha_{rs} > 0\,. \tag{2.14}$$

The exponential model is the simplest parametric hazard specification, which does not allow for time-dependent modelling. The Weibull and Gompertz specifications are useful to model monotonic upward or downward trends over time.

Whilst it is straightforward to specify time-dependent models for the hazards, calculating the transition probabilities of time-dependent multi-state processes are complicated. The forward and backwards equations in (2.7) and (2.8), respectively, are derived for processes for which $\mathbf{Q}$ is constant. For time-dependent multi-state processes, those equations can be extended as follows. Define a family of matrices $\mathbf{P}(t,u)$, for $u > t$, with elements $P(Y(u) = j|Y(t) = i)$. The forward and backward equations are given by

$$\frac{\partial \mathbf{P}(t,u)}{\partial u} = \mathbf{P}(t,u)\mathbf{Q}(u), \tag{2.15}$$

$$-\frac{\partial \mathbf{P}(t,u)}{\partial t} = \mathbf{Q}(t)\mathbf{P}(t,u), \tag{2.16}$$

respectively. Equations (2.15) and (2.16) are called the *Kolmogorov differential equations*. Finding a solution or approximated solution for these equation is crucial for practical multi-state modelling. Titman (2011) uses a numerical approximation to solve these differential equations. In this thesis, time-dependency is approached by using a piecewise-constant approximation to the hazards.

2.3 Piecewise-constant hazards approximation

This section describes a piecewise-constant approximation to take into account time-dependency of the hazards. Given a specified time-interval $(t_1, t_2]$, we aim to find an approximate value for the transition probability $\mathbf{P}(t_1, t_2)$, which is then employed in estimation. We next present two methods that can be used within the framework that will be developed in this thesis.

The first method uses the follow-up times in the data to define the grid for the piecewise-constant approximation for the individual likelihood contributions. In this case, for a given time-interval $(t_1, t_2]$ the transition probability $\mathbf{P}(t_1, t_2)$ is estimated by $\exp((t_2 - t_1)\mathbf{Q}(t_1))$. This approximation performs well in estimation depending on the volatility of the process and on the study design. For more volatile processes, the follow-up times have to be more frequent to capture the risk changes over time.

Alternatively, it is possible to impose a fixed grid to the piecewise-constant approximation as described in Van den Hout and Matthews (2008a). For a given time interval $(t_1, t_2]$, the transition probability $\mathbf{P}(t_1, t_2)$ can be calculated by imposing a grid for the piecewise-constant approximation, which is the same for all individual likelihood contributions. In this case, time intervals in the data are embedded in the grid. For example, say the grid is defined by $u_1, ..., u_M$. For the observed time interval $(t_1, t_2]$, determine j_1 and j_2 such that $u_{j_1} < t_1 \leq u_{j_1+1}$ and $u_{j_2} < t_2 \leq u_{j_2+1}$. The transition matrix for $(t_1, t_2]$ is then defined by

$$\mathbf{P}(t_1, t_2) = \mathbf{P}\left(t_1, u_{j_1+1}\right) \mathbf{P}\left(u_{j_1+1}, u_{j_1+2}\right) \times \cdots \times \mathbf{P}\left(u_{j_2}, t_2\right),$$

using generator matrices $\mathbf{Q}(u_{j_1}), \mathbf{Q}(u_{j_1+1}), \ldots, \mathbf{Q}(u_{j_2})$, respectively.

Van den Hout (2017) compares the performance of the piecewise-constant approximation and the exact method that solves the forward and backward differential equations as in (2.15) and (2.16). The simulation is carried out for the illness-death model with Gompertz hazards. The grid for the piecewise-constant approximation is defined by the data. Two study designs are investigated: states are observed yearly

and at the points 0, 4, 8, 9, 10, 11, 12. For the first case, the results of both methods are very similar. For the second case, the piecewise-constant approximation led to slightly biased results for the scale parameter, but similar results for the shape parameter. It is expected given the 4-year interval times. Therefore, the piecewise-constant approximation leads to satisfactory results as long as the time between observations is not too long relative to the volatility of the multi-state process.

For multi-state processes in which the sampling times are covariate dependent (e.g. individuals with a pre-condition are sampled more often), the approximation bias may lead to bias of the covariate effect. For such cases, imposing a grid to the piecewise-constant approximation can improve the estimation. Also, it is possible to include an interval-censored state in between two distant observation times (Van den Hout, 2017).

2.4 Maximum likelihood estimation

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$ be the parameters vector of a given multi-state model. These parameters can be estimated using the method of maximum likelihood. We first obtain the likelihood function of the sample data. Then we illustrate a scoring algorithm, which is used to obtain the maximum likelihood estimates of the model parameters.

2.4.1 Likelihood function of the model

Given a multi-state model, maximum likelihood inference can be used to estimate model parameters. In the presence of interval censoring, the likelihood function is constructed using transition probabilities.

Let the state space be $\mathscr{S} = \{1, 2, .., D\}$, with D the dead state. Consider a series of states $Y_1, ..., Y_n$ 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

$$P\left(Y_n = y_n, ..., Y_2 = y_2 | Y_1 = y_1, \boldsymbol{\theta}, \mathbf{t}, \mathbf{X}\right), \tag{2.17}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)^\top$ 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 first-order Markov assumption is used to define the distribution (2.17) of $Y_2, ..., Y_n$ 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 $\mathbf{X}$.

Next consider an individual i with observed values $y_1, ..., y_{n-1} \in \mathscr{S} \backslash D$, and a last observation y_n which is either a value in $\mathscr{S}$ or a code for right-censoring. The likelihood contribution for this individual is $L_i = \prod_{j=2}^{n} L_{ij}$, where

$$L_{ij} = \begin{cases} P\left(Y_j = y_j | Y_{j-1} = y_{j-1}, \boldsymbol{\theta}, t_{j-1}, \mathbf{x}_{j-1}\right) & \text{for} \quad j = 2, ..., n-1 \\ C(y_n | y_{n-1}) & \text{for} \quad j = n\,. \end{cases} \tag{2.18}$$

If a living state at t_n is observed, then $C(y_n|y_{n-1}) = P(Y_n = y_n | Y_{n-1} = y_{n-1})$, where part of the conditioning is ignored in the notation. If the state is right censored at t_n, then $C(y_n|y_{n-1}) = \sum_{s=1}^{D-1} P(Y_n = s | Y_{n-1} = y_{n-1})$. If the state at t_n is D, then known time of death is taken into account by defining

$$C(y_n|y_{n-1}) = \sum_{s=1}^{D-1} P(Y_n = s | Y_{n-1} = y_{n-1})\, q_{sD}(t_n). \tag{2.19}$$

Given N individuals, the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log L_i = \sum_{i=1}^{N} \sum_{j=2}^{n_i} \log L_{ij}, \tag{2.20}$$

where n_i is the number of observation times for individual i, which can include a right-censored observation.

The above definition of the likelihood function can also be found in Jackson (2011). Including time-dependency, as defined by the models in Section 2.2, does not affect the basic structure of the likelihood function. Similar expressions of the likelihood function can be found in Kalbfleisch and Lawless (1985), Kay (1986),

Gentleman et al. (1994), and Van den Hout (2017).

The log-likelihood function in (2.20) can be maximised by using a general-purpose optimiser or a scoring algorithm. The latter is illustrated in the next section and will be used for the analysis in this chapter.

2.4.2 Scoring algorithm

The maximum likelihood estimates of the vector of parameters $\boldsymbol{\theta}$ in the multi-state model (2.11) can be found by maximising the log-likelihood function (2.20) using numerical methods. This maximisation can be achieved using a scoring algorithm (Van den Hout, 2017), and it is described below.

Kalbfleisch and Lawless (1985) proposed a scoring algorithm for maximising the log-likelihood function of time homogeneous multi-state models. Given a piecewise-constant approximation to the time-dependency in the hazard model (2.11), their formulae can be used to maximise the likelihood function (2.20). In particular, those formulae are applied to the constituent intervals with constant hazards in the likelihood function. Notice that, in this case, we obtain an approximation to the log-likelihood function in (2.20), which is still the logarithm of a likelihood function.

As in Section 2.4.1, let $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)^\top$ be the vector of model parameters. For a given time interval $(t_1, t_2]$, the first-order derivatives of the transition probability matrix, $\partial \mathbf{P}(t_1, t_2)/\partial \theta_k$, is given in terms of $\partial \mathbf{Q}(t)/\partial \theta_k$. The latter is straightforward to derive for most parametric models, such as, Weibull and Gompertz hazard models. The likelihood contributions for exact death times and right-censoring are straightforward to deal with, as they are made up of transition probabilities.

To specify the scoring algorithm, the derivative of a transition matrix is presented first. Given a piecewise-constant approximation to the hazards, 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 $\mathbf{b} = (b_1, ..., b_D)^\top$, 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

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

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 $\mathbf{G}^{(k)} = \mathbf{A}\partial\mathbf{Q}/\partial\theta_k\mathbf{A}^{-1}$.

For the parametric time-dependent hazard models in Section 2.2, matrix $\partial\mathbf{Q}/\partial\theta_k$ is straightforward to derive.

The scoring algorithm can now be defined as follows. Let $\mathbf{g}(\boldsymbol{\theta})$ be the $q \times 1$ vector of first-order derivatives of the log-likelihood function in (2.20). This quantity is called the *gradient vector*. The kth entry of $\mathbf{g}(\boldsymbol{\theta})$ is given by

$$\sum_{i=1}^{N}\sum_{j=2}^{n_i}\frac{\partial}{\partial\theta_k}\log L_{ij}\,. \tag{2.21}$$

Let $\mathscr{I}(\boldsymbol{\theta})$ be the $q \times q$ matrix of expected negative second-order derivatives of the log-likelihood. This quantity is given by

$$\mathscr{I}(\boldsymbol{\theta}) = I\!E\left[-\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]. \tag{2.22}$$

The matrix $\mathscr{I}(\boldsymbol{\theta})$ is called the *Fisher information matrix*. This matrix can be approximated by defining the $q \times q$ matrix $\mathbf{M}(\boldsymbol{\theta})$ with (k,l) entry

$$\sum_{i=1}^{N}\sum_{j=2}^{n_i}\frac{\partial}{\partial\theta_k}\log L_{ij}\frac{\partial}{\partial\theta_l}\log L_{ij}\,. \tag{2.23}$$

The scoring algorithm provides that an estimate of the vector $\boldsymbol{\theta}$ at the $(v+1)$th cycle of the iterative procedure, $\boldsymbol{\theta}^{(v+1)}$, is

$$\boldsymbol{\theta}^{(v+1)} = \boldsymbol{\theta}^{(v)} + \mathbf{M}\big(\boldsymbol{\theta}^{(v)}\big)^{-1}\mathbf{g}\big(\boldsymbol{\theta}^{(v)}\big)\,.$$

for $v = 1,2,3\ldots$, where $\mathbf{g}\big(\boldsymbol{\theta}^{(v)}\big)$ is the gradient vector and $\mathbf{M}\big(\boldsymbol{\theta}^{(v)}\big)^{-1}$ is the inverse of the (approximated) Fisher information matrix, both evaluated at $\boldsymbol{\theta}^{(v)}$. The process iterates until the relative differences in the values of the parameter estimates satisfies $\max_{1\leq k\leq q}|\theta^{(v+1)} - \theta^{(v)}| < \delta$ for a suitable small positive value.

The asymptotic covariance matrix of the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is equal to the inverse of the Fisher information matrix $\mathscr{I}(\boldsymbol{\theta})^{-1}$, which can be approximated by $\mathscr{I}(\widehat{\boldsymbol{\theta}})^{-1}$. Hence, after convergence, the covariance matrix of the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is estimated by $\mathbf{M}(\widehat{\boldsymbol{\theta}})^{-1}$ (Van den Hout, 2017). The standard error of $\widehat{\theta_i} = (\widehat{\boldsymbol{\theta}})_i$ is given by

$$SE(\widehat{\theta_i}) \approx (\mathbf{M}(\boldsymbol{\theta})^{-1})_{ii}^{1/2}, \tag{2.24}$$

where $i = 1,\ldots,q$. The algorithm is implemented by the author in R in such a way that it is easy to vary transition-specific choices for parametric shapes.

2.5 Confidence intervals

The distribution of the maximum likelihood estimator can be used to construct confidence intervals for the estimate $\widehat{\boldsymbol{\theta}}$ and functions of them, such as the hazards and probability matrix (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 $(1-\varsigma/2)$, quantiles.

The parameter ς is usually set to 0.05. In this thesis, we approximate the covariance matrix $\mathbf{V}_{\boldsymbol{\theta}}$ by the inverse of the matrix $\mathbf{M}$ as defined in (2.23).

Sampling from a K-variate normal distribution $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is possible by using the Cholesky decomposition $\boldsymbol{\Sigma}=\mathbf{L}\mathbf{L}^{\top}$. First K draws are taken independently from the standard normal and collected in the $K\times 1$ vector $\mathbf{z}$. A multivariate draw from $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is then given by $\boldsymbol{\mu}+\mathbf{L}\mathbf{z}$.

2.6 Model selection

In multi-state models various hazards specifications can be used. Model selection is commonly used to select the best model among them. Model fitting can always be improved by adding more parameters. However, parsimonious models are easier to estimate and better for prediction. Model selection methods aim to find the balance between goodness-of-fit and parsimony. Ruppert et al. (2003) and Wood (2006) provide descriptions of several methods for model selection. Their method can be used to compare nested multi-state models, i.e., models with the same structure, but with different specifications. This thesis uses the Akaike Information Criterion (Akaike, 1998). This quantity is also know as AIC. This section follows the description presented in Ruppert et al. (2003) and Wood (2006).

The AIC selects models based on their fit to new data. The AIC is derived from the Kullback-Leibler (K-L) discrepancy. Suppose that the model density is $f_{\theta}(y)$, and that $f_0(y)$ is the true density, where θ denotes the parameter (or the vector of parameters) of f_{θ}. The K-L discrepancy is given by

$$K(f_{\theta},f_0)=\int\{\log[f_0(y)]-\log[f_{\theta}(y)]\}f_0(y)dy. \tag{2.25}$$

Equation (2.25) provides a measure of how well f_{θ} matches the truth. If $\widehat{\theta}$ is the maximum likelihood estimate of θ, then $K(f_{\theta},f_0)$ could be used to provide a measure of how well the the model f_{θ} is expected to fit a new set of data, not used to estimate $\widehat{\theta}$. It can be shown that $\mathbb{E}[K(f_{\theta},f_0)]\approx-\ell(\widehat{\theta})+q$, where $\ell(\widehat{\theta})$ is the log-likelihood function evaluated at the maximum likelihood estimate, and q represents

the number of model parameters. This measure can be used for model comparison. The Akaike Information Criterion is then defined as

$$AIC = -2\ell(\widehat{\theta}) + 2q. \tag{2.26}$$

Models with lower AIC values are considered to be closer to the true model, than models with higher AIC values. The term q penalises models with more parameters than necessary, counteracting the tendency of the likelihood to models with a large number of parameters.

2.7 Model validation

For multi-state models, the Markov assumption and parametric hazards specification are used to facilitate inference. Titman and Sharples (2010) reviewed diagnostics methods for multi-state models in the presence of interval censoring. We describe below the method that will be used in this thesis for model validation and model comparison.

A simple diagnostic method compares the model predictions of the entry time into an absorbing state with the Kaplan-Meier estimates (see Section 1.4). Suppose that all individuals start in the same state at baseline and progress to an absorbing state, and that the assumptions in the multi-state model are correct. Then, there should be close agreement between the empirical survival curve and the survival curve implied by the fitted multi-state model. In this case, survival time is in relation to the dead state. The pointwise 95% confidence intervals of the Kaplan-Meier are used as a benchmark to decide if any disagreement is within allowed bands. If the time scale is age, then age has to be transformed to time since beginning of the study. In this case, we can compare all individual survival curves. Notice that this method only checks one part of the model, the survival times from a specified state. However, it gives relevant information about model fitting. This method is used to assess model fit throughout this thesis.

Table 2.1: State table for the OCTO data for the history of State 3: number of times each pair of states was observed at successive observation times. The three living states are defined by grades of cognitive function

	To			
From	1	2	3	4
1	710	130	79	229
2	66	98	100	104
3	0	0	339	329

2.8 Applications

The methods presented in this chapter are illustrated with applications to the OCTO and CAV data sets. In what follows, model estimation is undertaken using the scoring algorithm presented in Section 2.4.2. Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)^\top$ be the vector with model parameters, where q depends on the application. The convergence criterion for the algorithm is to stop at iteration $v+1$ when $\max_{1\leq k\leq q} |\theta^{(v+1)} - \theta^{(v)}| < 10^{-6}$. For both data sets, models are specified with Gompertz hazards because this is a common assumption in applied research (Marioni et al., 2012; Robitaille et al., 2018) and can be used with the msm package.

2.8.1 Origins of variance in the oldest-old data

We fit a multi-model for the OCTO data defined as in Figure 1.11. Even though individuals in state 3 (severe cognitive impairment) can only move into state 4 (dead), some backward observations from state 3 are recorded. This is due to measurement error. It is possible to fit multi-state models with misclassification of states (Jackson et al., 2003). However, such an approach is out of the scope of this research. We define the OCTO data for the history of state 3, which means that once individuals move into state 3, they are only allowed to move into state 4. The resulting interval-censored multi-state process is summarised by the frequencies in Table 2.1.

Let t represent age minus 80. This is necessary to avoid numerical problems with the scoring algorithm. Because time of death is known, rather than being interval censored, the likelihood contribution of individuals observed in state $r < 4$ at time t and dead at time $t^* > t$ are given by $\sum_{s=1}^{3} P(Y(t^*) = s|Y(t) = r)\, q_{s4}(t^*)$. As

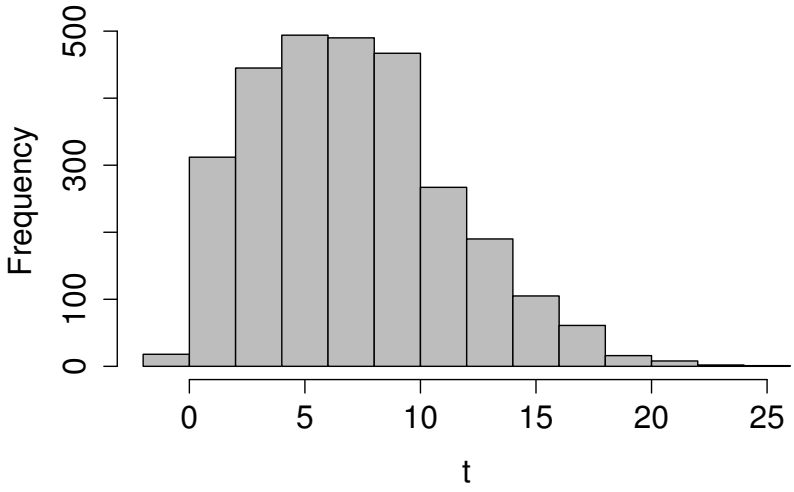

Figure 2.1: Histogram of age at baseline transformed by minus 80 in the OCTO data. The variable t represents age minus 80

Table 2.2: Parameter estimates for the four-state for the OCTO data. Estimated standard errors in parentheses. Time scale t is age in years minus 80

Intercept		*t*		*sex*	
$\alpha_{12.0}$	-2.157 (0.163)	$\alpha_{12.1}$	0.115 (0.024)	β_L	-0.325 (0.100)
$\alpha_{14.0}$	-3.185 (0.220)	$\alpha_{14.1}$	0.128 (0.032)	β_D	-0.334 (0.091)
$\alpha_{21.0}$	-1.383 (0.254)	$\alpha_{21.1}$	-0.020 (0.047)		
$\alpha_{23.0}$	-0.967 (0.162)	$\alpha_{23.1}$	0.056 (0.023)		
$\alpha_{24.0}$	-3.775 (0.719)	$\alpha_{24.1}$	0.177 (0.078)		
$\alpha_{34.0}$	-1.463 (0.130)	$\alpha_{34.1}$	0.060 (0.013)		

described in Section 2.3, transition probabilities for the likelihood function are calculated by using a piecewise-constant approximation to the hazards. For the OCTO data, the mean length of follow-up times is 1.994 years with standard deviation of 1.098 and median 1.986. Assuming that change in transition intensities in relation to the frequency of observation can be assessed in intervals of approximately 2 years, we can use the data to define the grid for the piecewise-constant approximation.

The proportional hazards model with Gompertz specification and dependence

on the covariate *sex* is given by

$$q_{rs}(t) = \exp\left(\alpha_{rs.0} + \alpha_{rs.1}t + \beta_{rs}sex\right), \tag{2.27}$$

where $(r,s) \in \{(1,2),(1,4),(2,1),(2,3),(2,4),(3,4)\}$ and *sex* is $0/1$ for men/women. For the transition between the living states, the constraints on the coefficients for *sex* are $\beta_{12} = \beta_{23} = \beta_L$, except for transition from 2 to 1 where $\beta_{21} = 0$. For the transitions into the dead state, the constraints are $\beta_{14} = \beta_{24} = \beta_{34} = \beta_D$. This model has a total of 14 parameters and $AIC = 5342.837$.

The estimated hazards (solid lines) for women and 95% confidence intervals (dashed lines) are presented in Figure 4.5. The confidence intervals are obtained by simulation with $b = 1000$ replications. The risks of moving across states are increasing throughout the length of the study, except for the transition from 2 to 1 which is decreasing. The confidence intervals are fairly wide after approximately 15 years due to the dependence on the slope parameter as t increases. Figure 2.1 illustrates the histogram of age at baseline minus 80.

Table 2.2 presents the parameter estimates for model (2.27) for the OCTO data. The effect of *sex* for the living states is very close to the effect of *sex* for the transition into the dead state, $\widehat{\beta}_L = -0.325$ and $\widehat{\beta}_D = -0.334$.

Figure 2.3 depicts the baseline-specific survival as estimated by the model and as described by the Kaplan-Meier curves. Individual survival curves (in grey) are shifted to the *years since baseline* so that we can compare them and their mean to the Kaplan-Meier curve. This is necessary because individuals have different ages at baseline. For survival given baseline states 2 and 3, there is some discrepancy between the model-based mean survival and the Kaplan-Meier curve, but overall the fit is reasonably good. Although this is not a proper goodness-of-fit test, the comparison shows that the model is able to capture the attrition due to death during the follow-up.

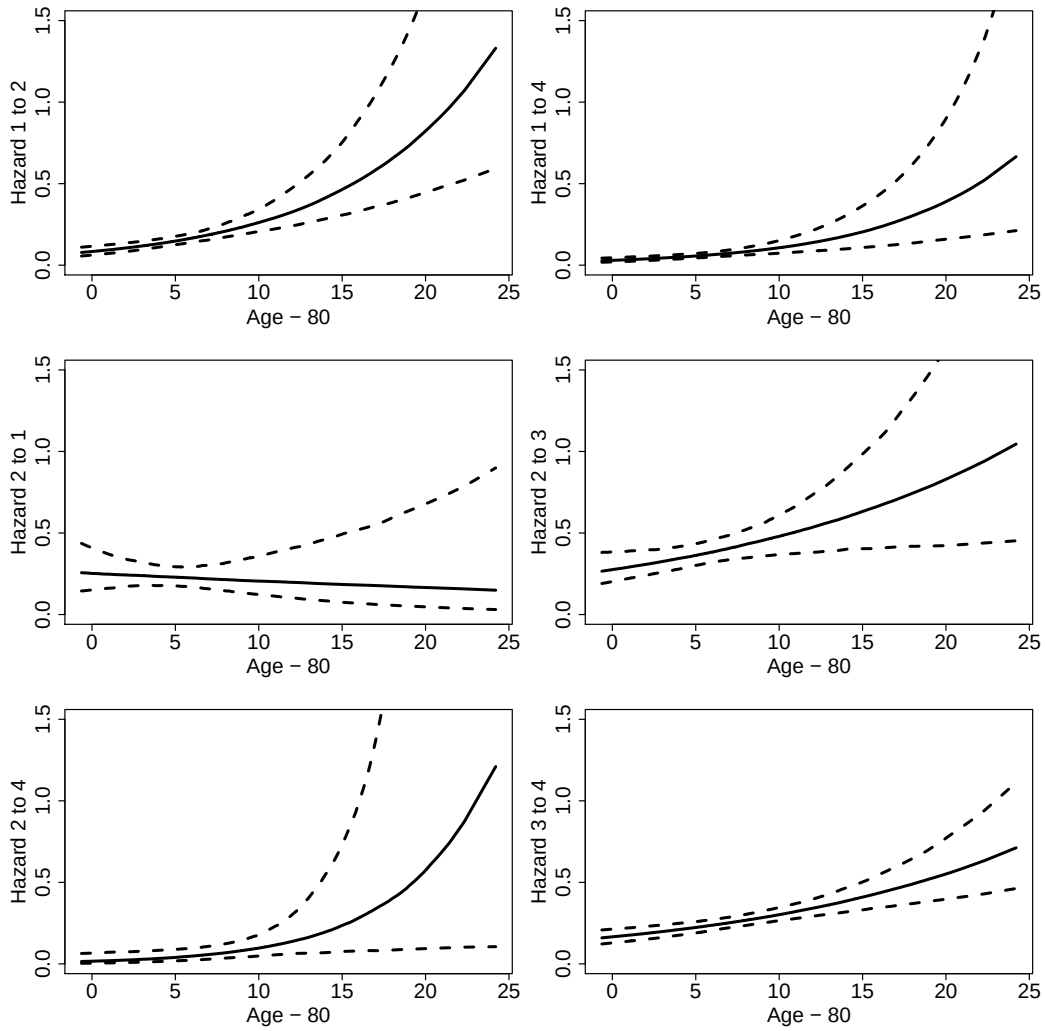

Figure 2.2: Estimated Gompertz hazards (solid lines) for women, with 95% confidence intervals (dashed lines). The confidence intervals are obtained by simulation with $b = 1000$ replications

2.8.2 Cardiac allograft vasculopathy data

For the CAV data as described in Section 1.6.1, some backward transitions are recorded from states 3 and 2. However, the process is biologically irreversible and of particular interest is the onset of CAV. In order to investigate this, an illness-death without recovery model can be defined. The states are classified as Healthy (1) if the patient has not developed the disease, CAV (2) if the patient has developed moderate or severe CAV and Dead (3) if the patient has died, see Figure 2.4. Also, diagnosis of ischaemic heart disease (IHD) and donor age are known to be major risk factors of disease onset (Titman, 2011). Only data until 15 years are considered, since after

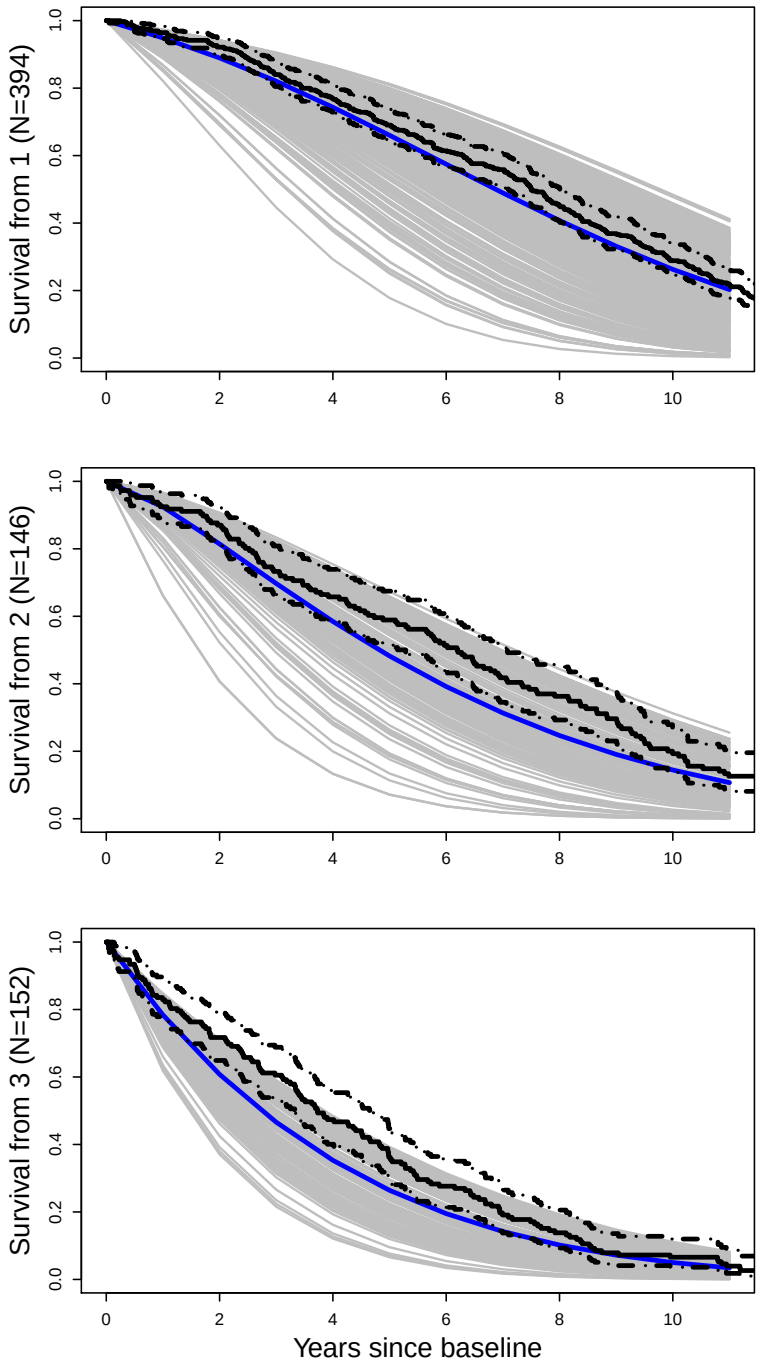

Figure 2.3: Comparison of model-based survival from states 1, 2, and 3 with Kaplan-Meier curves. Model-based survival: grey lines for individuals, blue lines for the mean of the individual survival curves. Kaplan-Meier in black lines with 95% confidence intervals. Frequencies for baseline state along vertical axes

this time data are scarce and we cannot estimate the model for times beyond that time. Titman (2011) used a similar formatting of the CAV data. Table 2.3 gives the number of times each pair of states was observed at successive observation times.

Table 2.3: State table for the CAV data: number of times each pair of states was observed at successive observation times. The two living states are defined by CAV severity

	To			
From	1	2	3	99
1	1336	224	139	252
2	0	412	110	105

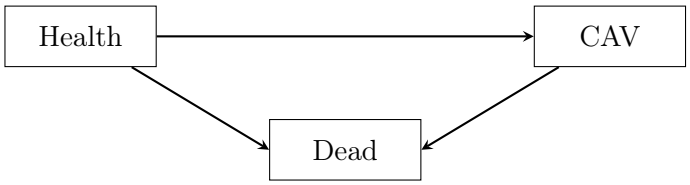

Figure 2.4: Illness-death without recovery model for disease progression after transplant for the CAV data

We fit a progressive three-state model for the CAV data defined as in Figure 2.4. 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 are dead at time $t^* > t$ are given by $\sum_{s=1}^{2} P(Y(t^*) = s|Y(t) = r)\, q_{s3}(t^*)$. As described in Section 2.3, transition probabilities for the likelihood function are calculated using a piecewise-constant approximation to the hazards. For the CAV data, the mean length of follow-up times is 1.622 years with standard deviation of 0.972 and median 1.258. Assuming that change in transition intensities in relation to the frequency of observation can be assessed in intervals of approximately one year, we can use the data to define the grid for the piecewise-constant approximation.

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

$$q_{rs}(t) = \exp\left(\alpha_{rs.0} + \alpha_{rs.1}t + \beta_1 dage + \beta_2 IHD\right), \tag{2.28}$$

where $(r,s) \in \{(1,2),(1,3),(2,3)\}$. This model has 8 parameters and $AIC = 3190.752$.

The estimated Gompertz hazards for subjects with IHD and donor age of 26

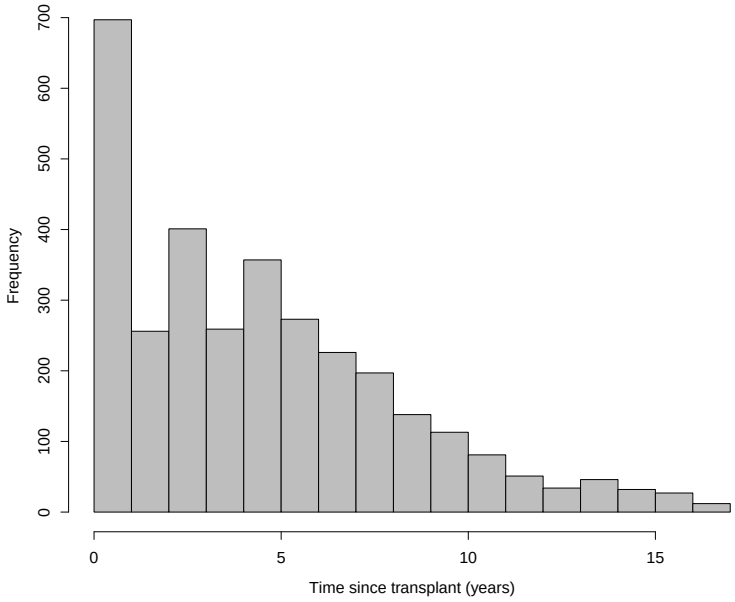

Figure 2.5: Histogram of time since transplant in the CAV data

Table 2.4: Parameter estimates for the illness-death without recovery for the CAV data. Estimated standard errors in parentheses. The variable t represents time since baseline

Intercept		t		*Covariates*	
$\alpha_{12.0}$	-3.166 (0.175)	$\alpha_{12.1}$	0.110 (0.022)	β_1	0.014 (0.004)
$\alpha_{13.0}$	-3.682 (0.204)	$\alpha_{13.1}$	-0.195 (0.064)	β_2	0.296 (0.089)
$\alpha_{23.0}$	-3.585 (0.268)	$\alpha_{23.1}$	0.101 (0.032)		

(solid lines) and 95% confidence intervals (dashed lines) are presented in Figure 2.6. The confidence intervals are calculated using simulation with $b = 1000$, as described in Section 2.5. The risks of moving from state 1 (Healthy) to state 2 (CAV), and from state 2 (CAV) to state 3 (Dead) increase over the years. The risk of going from state 1 to state 3 (Dead) is very low and decreasing over the years. Similarly to what happened in the application to the OCTO data (Section 2.8.1), the confidence intervals become wider towards the end of the study. The histogram of time since transplant in Figure 4.8 shows that data become scarce after approximately 10 years.

Table 2.4 illustrates the parameter estimates for model (2.28) for the CAV data. For the covariates effects, $\widehat{\beta_1} = 0.018$ and $\widehat{\beta_2} = 0.277$ indicating that donor age and IHD increase the risks of disease progression and death.

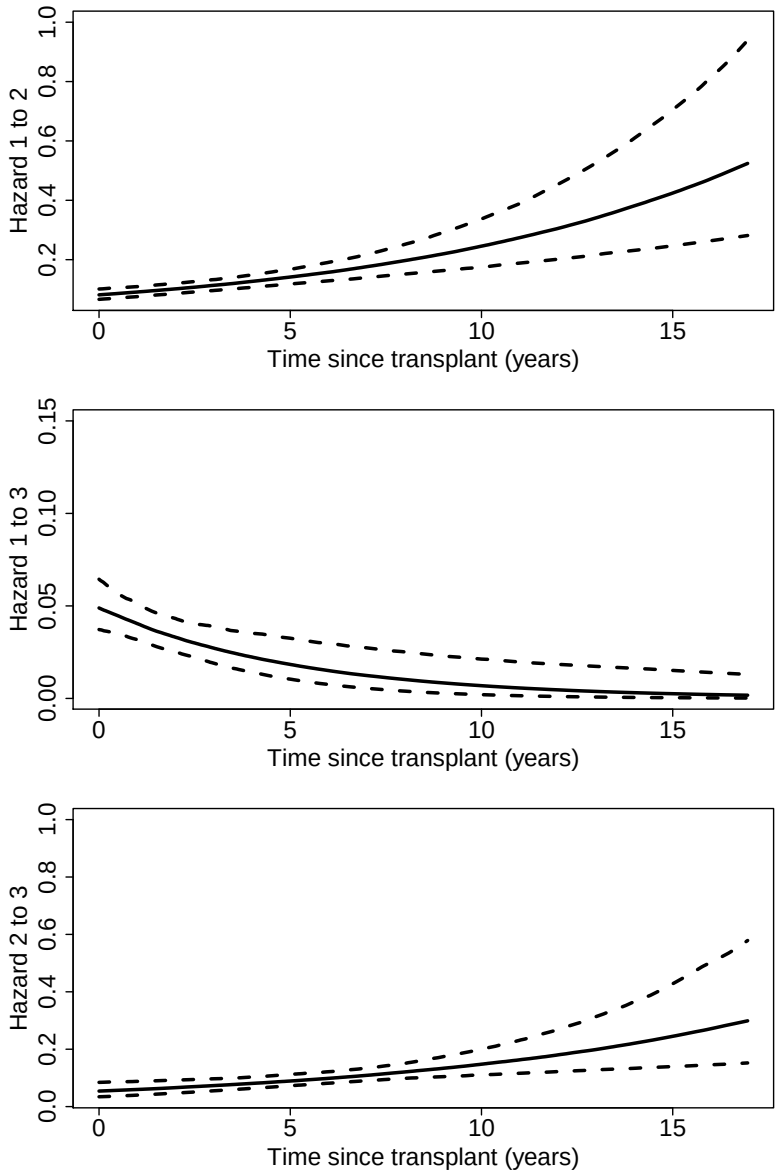

Figure 2.6: Estimated Gompertz hazards for subjects with IHD and with donor age of 26 (solid lines), with 95% confidence intervals (dashed lines). The confidence intervals are obtained by simulation with $b = 1000$ replications

Figure 2.7 shows baseline survival as estimated by the model and as described by the Kaplan-Meier curves. The model predicts the survival reasonably well up to ten years. However, there is some discrepancy between the estimate survival curve and the Kaplan-Meier curve after ten years, which is an indication of lack of fit.

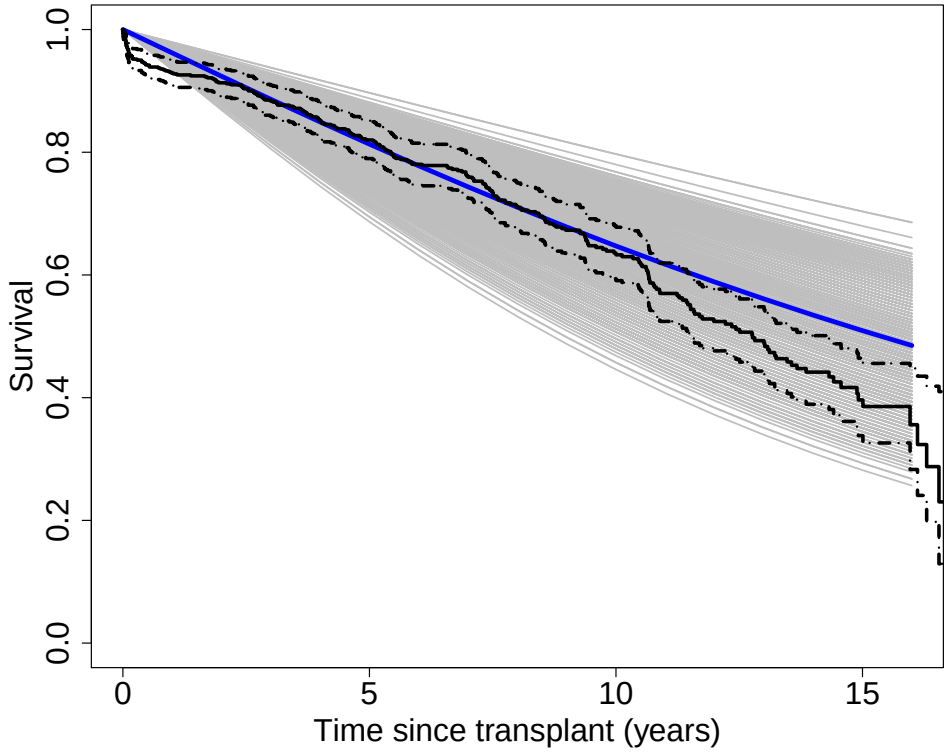

Figure 2.7: Comparison of model-based survival from states 1 with Kaplan-Meier curves. Model-based survival: grey lines for individuals, blue line for the mean of the individual survival curves. Kaplan-Meier in black lines with 95% confidence intervals

2.9 Discussion

This chapter presented the general formulation of parametric multi-state models for interval-censored data. We show in an application to the OCTO data that a Gompertz hazard specification results in a satisfactory model fit. In the application with the CAV data, we demonstrate that such parametric model assumption can be too restrictive. This thesis aims to investigate how model fit can be improved by using flexible hazards specification with splines, without defining a specific parametric form. Even though there are a wide range of more flexible model, deciding on a specific model can be difficult. The next chapter introduces the multi-state models with splines, which aim to overcome restrictive hazards models such as Gompertz and Weibull.

Chapter 3

Multi-state models with splines

This chapter begins with a brief introduction to smoothing methods. This will include a detailed description of cubic regression splines and *P*-splines basis functions. We then show how to specify transition-specific hazard functions with splines. A penalised maximum likelihood method is developed to estimate model parameters. The method uses a scoring algorithm to maximise the penalised log-likelihood function and a grid search to select the optimum amount of smoothing. The chapter then concludes with an application to the ELSA data, where parametric and semi-parametric transition-specific hazards specifications are explored to model the multi-state processes. A substantial part of the material in this chapter has been published in *Statistics in Medicine* as a paper entitled *Flexible multistate models for interval-censored data: Specification, estimation, and an application to ageing research* (Machado and van den Hout, 2018). We acknowledge that part of the analysis of the ELSA data is also in Van den Hout (2017).

3.1 Introduction

The application to the CAV data in Chapter 2 shows that parametric hazards specification can lead to poor model fitting. In this section, we present an introduction to smoothing methods, which are a general technique to estimate nonparametric functions. These methods allow for flexible modelling without making strong assumption about the functional forms underlying the data. We describe how to approach the simple case of estimating a univariate nonparametric function. The methods

are subsequently extended to specify and estimate multi-state models with splines. This sections follows the description presented in Ruppert et al. (2003) and Wood (2006).

3.1.1 Smoothing methods

Consider the problem of estimating a function g from a set of n data points (x_i, y_i) for $i = 1, \ldots, n$, where g is continuous on $[x_1, x_n] \in \mathbb{R}$ with absolutely continuous first derivatives. The nonparametric function f can be used as a predictor

$$y_i = f(x_i) + \varepsilon_i, \tag{3.1}$$

where ε_i are the error terms and $\mathbb{E}(\varepsilon_i) = 0$.

In practice, f is represented in terms of spline basis functions. Splines are curves made up of sections of polynomials joined together so that they are continuous in value, as well as in first and second derivatives. The points at which the sections join are known as the *knots* of the spline. The locations of the knots must be chosen. Typically the knots would either be evenly spaced through the range of the observed x values, or placed at quantiles of the distribution of unique x values.

The problem of estimating the nonparametric function f with splines can be formulated as follows. Let $\mathbf{B}(x) = [B_1(x), \ldots, B_q(x)]^\top$ be the vector of spline basis functions evaluated at x and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^\top$ the parameter vector so that the nonparametric function can be written as $f(x) = \boldsymbol{\alpha}^\top \mathbf{B}(x)$. A *penalised spline* is defined as $\widehat{\boldsymbol{\alpha}}^\top \mathbf{B}(x)$, where $\widehat{\boldsymbol{\alpha}}$ is the minimiser of

$$\sum_{i=1}^{n} \{y_i - \boldsymbol{\alpha}^\top \mathbf{B}(x_i)\}^2 + \lambda \boldsymbol{\alpha}^\top \mathbf{S} \boldsymbol{\alpha}, \tag{3.2}$$

for some symmetric positive semidefinite matrix $\mathbf{S}$ and scalar $\lambda > 0$. The matrix $\mathbf{S}$ is called the *penalty matrix*. The scalar λ is known as the *smoothing parameter*, which is used to control the function smoothness. Various definitions of spline basis functions can be found in Wood (2006). This thesis uses *cubic regression splines* and *P-splines*, which are introduced in Sections 3.1.2 and 3.1.3, respectively. The

smoothing method discussed in this section can be generalised to more complex settings such as generalised additive models (Wood, 2006).

3.1.2 Cubic regression splines

Cubic regression splines are parametrised in terms of the values of polynomial functions at the knots. In particular, the knots can be placed considering the percentiles of the distributions of unique x values. This means that more knots are placed where there is more data. This is appealing for multi-state modelling because multi-state processes commonly become scarce towards the end of study. If knots are placed where there is no data, models cannot be estimated.

Consider defining a cubic spline function, $f(x)$, with k knots, $x_1,\ldots,x_k$. Let $f(x_j)=\alpha_j$ and $f''(x_j)=\delta_j$ for $j=1,\ldots,k$. Then the spline can be written as

$$f(x)=a_j^-(x)\alpha_j+a_j^+(x)\alpha_{j+1}+c_j^-(x)\delta_j+c_j^+(x)\delta_{j+1}, \tag{3.3}$$

for the interval $[x_j,x_{j+1}]$. The basis function $a_j^-(x),a_j^+(x),c_j^-(x)$ and $c_j^+(x)$ are defined as

$$\begin{aligned}
a_j^-(x) &= (x_{j+1}-x)/h_j,\\
a_j^+(x) &= (x-x_j)/h_j,\\
c_j^-(x) &= [(x_{j+1}-x)^3/h_j-h_j(x_{j+1}-x)]/6,\\
c_j^+(x) &= [(x-x_j)^3/h_j-h_j(x-x_j)]/6,
\end{aligned}$$

where $h_j=x_{j+1}-x_j$. Define the matrices $\mathbf{G}$ and $\mathbf{D}$ as

$$\begin{aligned}
D_{i,i} &= 1/h_i,\\
D_{i,i+1} &= -1/h_i-1/h_{i+1},\\
D_{i,i+2} &= 1/h_{i+1},
\end{aligned}$$

for $i = 1, \ldots, k-2$, and,

$$\begin{aligned} G_{i,i+1} &= h_{i+1}/6, \\ G_{i+1,i} &= h_{i+1}/6, \end{aligned}$$

for $i = 1, \ldots, k-3$. The conditions that the spline must have continuous second derivatives, at x_j, and should have zero second derivative at x_1 and x_k, imply that

$$\mathbf{G}\boldsymbol{\delta}^{-} = \mathbf{D}\boldsymbol{\alpha}, \tag{3.4}$$

where $\boldsymbol{\delta}^{-} = (\delta_2, \ldots, \delta_{k-1})^{\top}$. Defining $\mathbf{F}^{-} = \mathbf{G}^{-1}\mathbf{D}$, and

$$\mathbf{F} = \begin{bmatrix} \mathbf{0} \\ \mathbf{F}^{-} \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{0}$ is a row of zeros, we have that $\boldsymbol{\delta} = \mathbf{F}\boldsymbol{\alpha}$. Hence, the spline can be re-written in terms of $\boldsymbol{\alpha}$ as

$$f(x) = a_j^{-}(x)\alpha_j + a_j^{+}(x)\alpha_{j+1} + c_j^{-}(x)\mathbf{F}_j\boldsymbol{\alpha} + c_j^{+}(x)\mathbf{F}_j\boldsymbol{\alpha}, \tag{3.5}$$

for $x_j \leq x \leq x_{j+1}$. Equation (3.5) can be further written as

$$f(x) = \sum_{i=1}^{k} \alpha_i B_i(x), \tag{3.6}$$

where $B_i(x)$ are given implicitly by (3.5). Therefore, given a set of x values, at which to evaluate the spline, it is possible to obtain a model matrix mapping $\boldsymbol{\alpha}$ to evaluate the spline. It can be shown that

$$\int_{x_1}^{x_k} f''(x)^2 dx = \boldsymbol{\alpha}^{\top}\mathbf{D}^{\top}\mathbf{G}^{-1}\mathbf{D}\boldsymbol{\alpha}, \tag{3.7}$$

that is, $\mathbf{S} = \mathbf{D}^{\top}\mathbf{G}^{-1}\mathbf{D}$ is the penalty matrix (Wood, 2000) . Figure 3.1 illustrates

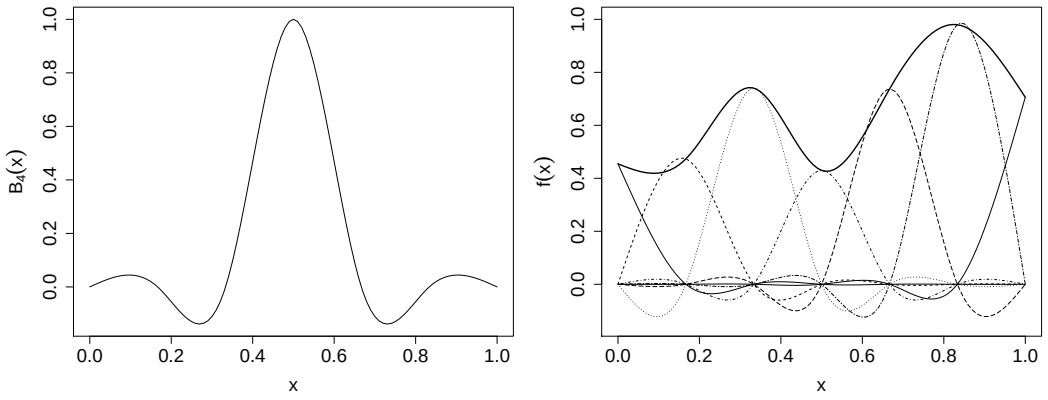

Figure 3.1: The left hand panel illustrates one basis function, $B_4(x)$, for a cubic regression spline. On the right hand panel, the various curves of medium thickness show the basis functions, $B_i(x)$, of a cubic regression spline, each multiplied by its coefficients α_j. These scaled basis functions are summed to get the smooth curve illustrated by the thick continuous curve

how a smooth function can be represented in terms of cubic regression spline basis functions. The cubic regression splines as defined here are implemented in the package mgcv in R (Wood, 2007).

3.1.3 *P*-splines

Simpler cubic splines can be represented by a *B*-splines basis. The *B*-splines basis functions are appealing because these functions are strictly local. Each basis function is only non-zero over the intervals between $m+3$ adjacent knots, where m is the order of the basis (e.g., $m = 2$ for cubic splines). The $k+m+1$ knots, $x_1 < x_2 < \ldots < x_{k+m+1}$, define a k parameter *B*-spline basis, where the interval over which the spline is to be evaluated lies within $[x_{m+2}, x_k]$. An $(m+1)^{th}$ order spline can be represented as

$$f(x) = \sum_{i=1}^{k} B_i^m(x)\alpha_i, \tag{3.8}$$

where the *B*-spline basis functions are most conveniently defined recursively as follows:

$$B_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m+2} - x_{i+1}} B_{i+1}^{m-1}(x),$$

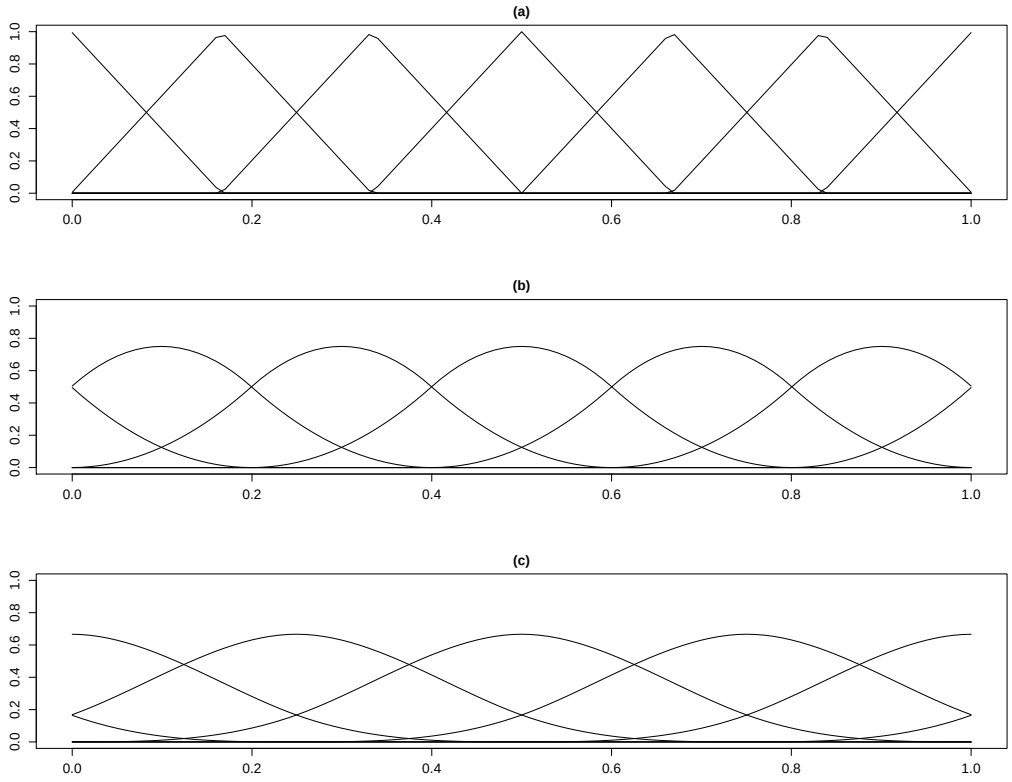

Figure 3.2: Illustrations of *B*-splines basis functions of degrees (a) one, (b) two, and (c) three

for $i = 1, \ldots, k$, and

$$B_i^{-1}(x) = \begin{cases} 1 & x_i \leq x < x_{i+1} \\ 0 & \text{otherwise} \end{cases}.$$

Figure 3.2 shows the *B*-splines basis of degree 1, 2 and 3 for the case of seven evenly spaced knots. Figure 3.3 illustrates how a smooth curve can be represented in terms of *B*-splines basis functions of degree two and three.

P-splines are a smoother using *B*-splines basis functions, defined on evenly spaced knots, with a difference penalty applied directly to the parameters, α_i, to control function smoothness. To illustrate how this works, suppose we decide to penalise the squared difference between adjacent α_i values. Then the penalty would be

$$\mathscr{P} = \sum_{i=1}^{k-1} (\alpha_{i+1} - \alpha_i)^2 = \alpha_1^2 - 2\alpha_1\alpha_2 + 2\alpha_2^2 - 2\alpha_2\alpha_3 + \ldots + \alpha_k^2,$$

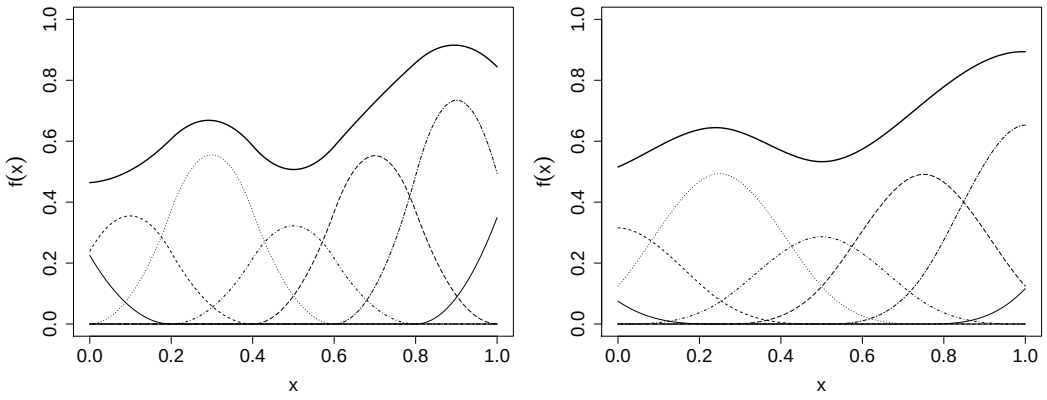

Figure 3.3: Illustration of smooth curves made up of *B*-spline basis functions. On the left hand panel, the dashed curves show *B*-splines basis functions with $m = 1$ multiplied by their associated coefficients. The thick solid smooth curve is the sum of the scaled basis functions. The right hand panel shows the same, but for *B*-splines basis function with $m = 2$

which can be written in matrix form as

$$\mathscr{P} = \boldsymbol{\alpha}^{\top} \begin{bmatrix} 1 & -1 & 0 & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \boldsymbol{\alpha}. \tag{3.9}$$

Equation 3.9 can be written as $\mathscr{P} = \boldsymbol{\alpha}^{\top}\mathbf{D}^{\top}\mathbf{D}\boldsymbol{\alpha}$, where the matrix $\mathbf{D}$ is the *difference operator* matrix (Eilers and Marx, 1996). *P*-splines are straightforward to set up and use, and allow for good flexibility as any order of penalty can be combined with any order of *B*-spline basis. Their main disadvantage is that the simplicity is somewhat diminished if uneven knot spacing is required (Wood, 2006).

3.2 Model representation

We showed in Section 2.2 that it is straightforward to define parametric time-dependent multi-state models. The specification of multi-state models with splines can be done in a similar way, as spline models can be seen as a type of parametric models.

Recall that a time-dependent hazard regression model for transition intensities combines baseline hazards with log-linear regression. Time-dependent models can be defined by using proportional hazards model for transition r to s, $r \neq s$, as follows

$$q_{rs}(t) = q_{rs.0}(t) \exp\left(\boldsymbol{\beta}_{rs}^{\top} \mathbf{x}\right), \tag{3.10}$$

where $q_{rs.0}(t)$ is the baseline hazard function, $\mathbf{x} = (x_1, \ldots, x_p)^{\top}$ is a covariate vector and $\boldsymbol{\beta}_{rs} = (\beta_{rs.1}, \ldots, \beta_{rs.p})^{\top}$ is vector of unknown parameters. In Chapter 2, examples of parametric shapes for $q_{rs.0}(t)$ are presented. We next describe the non-parametric specification of $q_{rs.0}(t)$ with splines. Each baseline hazard can be approximated by the exponential of a linear combination of K_{rs} spline base functions $B_k(t)$ and regression coefficients $\alpha_{rs.k} \in \mathbb{R}$ as follows

$$q_{rs.0}(t) = \exp\left(\sum_{k=1}^{K_{rs}} \alpha_{rs.k} B_k(t)\right). \tag{3.11}$$

Let the number of spline basis functions be large. 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. 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). With relation to the number of knots, $K_{rs} = 10$ is usually enough and larger number of splines basis functions will not change the estimated functions.

In this chapter, we focus on the use of P-splines for the basis functions $B_k(t)$. However, the method is implemented in a way that is easy to employ other spline definitions and corresponding penalties. Flexible multi-state models can be defined by P-splines and a combination of parametric hazards specification presented in Section 2.2. Applications of P-splines to multi-state models can also be found in

Kneib and Hennerfeind (2008).

3.3 Penalised maximum likelihood estimation

Multi-state models with splines can be fitted to a set of longitudinal data using penalised maximum likelihood estimation. In this section, we present the penalised log-likelihood function, and a scoring algorithm for finding the penalised likelihood estimates of the model parameters.

3.3.1 Penalised log-likelihood function

For semi-parametric multi-state models, at least one baseline hazard function is specified with *P*-splines. If the model for a transition is specified with *P*-splines, we define a large set of equidistant knots. To control the smoothness of the estimated curve, a penalty based on finite differences of the coefficient of adjacent *P*-splines is imposed on the log-likelihood function. Let $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ represent the vector of parameters associated to the parametric, non-parametric and covariates components of a multi-state model, respectively. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\xi}^\top)^\top$ be the full set of parameters and $\ell(\boldsymbol{\theta})$ be the log-likelihood function of a semi-parametric multi-state model. This function is constructed as described in Section 2.4.1. The penalised log-likelihood function is given by

$$\begin{aligned} \ell_p(\boldsymbol{\theta}) &= \ell(\boldsymbol{\theta}) - \frac{1}{2}\sum_{j=1}^{s} \lambda_j \boldsymbol{\alpha}_j^\top \mathbf{D}_j^\top \mathbf{D}_j \boldsymbol{\alpha}_j \\ &= \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\theta}, \end{aligned} \tag{3.12}$$

where s is defined as the number of transitions with *P*-splines, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jK})^\top$ is assumed to have the same dimension for each hazard approached with splines, $\mathbf{D}_j$ is the matrix representation of the difference operator of adjacent *P*-splines, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_s)^\top$ is the vector of smoothing parameters and $\mathbf{S}_{\boldsymbol{\lambda}}$ is the penalty matrix. $\mathbf{S}_{\boldsymbol{\lambda}}$ is a block diagonal matrix with blocks $\lambda_j \mathbf{D}_j^\top \mathbf{D}_j$ for penalising *P*-splines parameters and zeros elsewhere (Gray, 1992).

3.3.2 Parameter estimation

The penalised log-likelihood function (3.12) can be maximised using a numerical method that consists of two parts. First, given a grid of values for the smoothing parameters vector, we aim to find an estimate of the model parameters for each vector in the grid. Second, we select the smoothing parameters vector that minimises an information criterion such as the AIC. We next describe how to perform the first part of the method, that is, how to maximise the penalised log-likelihood function in (3.12) for fixed smoothing parameters vector.

Given a piecewise-constant approximation to the time-dependency in the hazard model (3.10), a scoring algorithm can be used to maximise the penalised log-likelihood function (3.12) for a fixed value of the smoothing parameter vector $\boldsymbol{\lambda}$. In particular, the scoring algorithm presented in Section 2.4.2 can be extended and used to find the penalised maximum likelihood estimates.

Let $\mathbf{g}_p(\boldsymbol{\theta})$ be the $q \times 1$ vector of first-order derivatives of the penalised log-likelihood function in (3.12). This quantity is called the *penalised gradient* and is given by

$$\begin{aligned} \mathbf{g}_p(\boldsymbol{\theta}) &= \partial \ell_p(\boldsymbol{\theta})/\partial \boldsymbol{\theta} && (3.13)\\ &= \partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta} - \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\theta} && (3.14)\\ &= \mathbf{g}(\boldsymbol{\theta}) - \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\theta} && (3.15) \end{aligned}$$

where $\mathbf{g}(\boldsymbol{\theta})$ is the gradient vector as defined in (2.21). Let $\mathscr{I}_p(\boldsymbol{\theta})$ be the $q \times q$ matrix of the expected negative second-order derivatives of the log-likelihood function. This matrix is known as *penalised Fisher information* and is given by

$$\begin{aligned} \mathscr{I}_p(\boldsymbol{\theta}) &= I\!E \left[-\partial^2 \ell_p(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top \right] && (3.16)\\ &= I\!E \left[-\partial^2 \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top + \mathbf{S}_{\boldsymbol{\lambda}} \right] && (3.17)\\ &= \mathscr{I} + \mathbf{S}_{\boldsymbol{\lambda}}, && (3.18) \end{aligned}$$

where $\mathscr{I}(\boldsymbol{\theta})$ is the Fisher information matrix as defined in (2.22).

The scoring algorithm provides that an estimate of the vector $\boldsymbol{\theta}$ at the $(v+1)$th cycle of the iterative procedure, $\boldsymbol{\theta}^{(v+1)}$, is

$$\boldsymbol{\theta}^{(v+1)} = \boldsymbol{\theta}^{(v)} + \mathscr{I}_p\big(\boldsymbol{\theta}^{(v)}\big)^{-1}\mathbf{g}_p\big(\boldsymbol{\theta}^{(v)}\big)\,. \tag{3.19}$$

for $v = 1,2,3\ldots$, where $\mathbf{g}_p\big(\boldsymbol{\theta}^{(v)}\big)$ is the penalised gradient vector and $\mathscr{I}_p\big(\boldsymbol{\theta}^{(v)}\big)^{-1}$ is the inverse of the penalised Fisher information matrix, both evaluated at $\boldsymbol{\theta}^{(v)}$. The process iterates until the relative differences in the values of the parameter estimates satisfies $\sum_{k=1}^{q}|\theta_k^{(v)} - \theta_k^{(v+1)}| < \delta$ for a suitable small positive value δ.

The asymptotic covariance matrix of the penalised maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is given by the inverse of the penalised Fisher information matrix $\mathscr{I}_p(\boldsymbol{\theta})^{-1}$ (Gray, 1992), which can be approximated by $\mathscr{I}_p(\widehat{\boldsymbol{\theta}})^{-1}$. Thus, the standard error of $\widehat{\theta}_i = (\widehat{\boldsymbol{\theta}})_i$ is given by

$$SE(\widehat{\theta}_i) \approx (\mathscr{I}_p(\widehat{\boldsymbol{\theta}})^{-1})_{ii}^{1/2}, \tag{3.20}$$

for $i = 1,\ldots,q$.

As discussed in Section 2.4.2, it is not possible to calculate the Fisher information matrix, and it is approximated by the $q \times q$ matrix $\mathbf{M}(\boldsymbol{\theta})$ with (k,l) entry

$$\sum_{i=1}^{N}\sum_{j=2}^{n_i}\frac{\partial}{\partial\theta_k}\log L_{ij}\frac{\partial}{\partial\theta_l}\log L_{ij}. \tag{3.21}$$

Hence, the penalised Fisher information matrix in (3.18) is approximated by

$$\mathbf{M}_p(\boldsymbol{\theta}) = \mathbf{M}(\boldsymbol{\theta}) + \mathbf{S}_{\boldsymbol{\lambda}}\,. \tag{3.22}$$

The matrix $\mathscr{I}_p$ is then replaced by $\mathbf{M}_p$ in equations (3.19) and (3.20).

The algorithm is implemented by the author in R in such a way that it is straightforward to vary transition-specific choices for parametric shapes. An example of such a model is explored in the application, where Gompertz and spline models are used to specify the hazards.

3.3.3 Smoothing parameter estimation

Estimating the optimal value for the smoothing parameters $\boldsymbol{\lambda}$ is crucial for fitting models with splines (Gu and Wahba, 1991). Given a grid of smoothing parameter vectors, the optimum value can be defined as the one with the smallest AIC. The AIC definition presented in Section 2.6 can be extended for semi-parametric models as follows,

$$\text{AIC}(\boldsymbol{\lambda}) = -2\ell_p(\boldsymbol{\theta}) + 2df, \tag{3.23}$$

where df is a measure of model complexity, and $\ell_p(\boldsymbol{\theta})$ is the penalised log-likelihood function. The quantity df is called the *degrees of freedom*. For parametric models, the degrees of freedom are equal to the number of independent parameters in the model. For semi-parametric models with splines, the degrees of freedom can be defined as

$$df(\boldsymbol{\lambda}) = \text{tr}[\mathscr{I}(\mathscr{I} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}], \tag{3.24}$$

where $\mathscr{I}$ is the Fisher information matrix and $\mathbf{S}_{\boldsymbol{\lambda}}$ is the penalty function (Gray, 1992). A similar definition of degrees of freedom is given in Commenges et al. (2007). If the vector of smoothing parameters is zero, $\boldsymbol{\lambda} = \mathbf{0}$, there is no penalty and the degrees of freedom is given by

$$\begin{aligned} df(\boldsymbol{\lambda}) &= \text{tr}[\mathscr{I}(\mathscr{I} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}] \\ &= \text{tr}[\mathscr{I}(\mathscr{I})^{-1}] \\ &= \text{tr}[\boldsymbol{I}] = q, \end{aligned}$$

where q is the number of parameters in the model, and $\boldsymbol{I}$ is the $q \times q$ identity matrix. Therefore, for $df = q$ the AIC definition in (3.23) coincides with the definition provided in Section 2.6. If the smoothing parameters vector is not zero, then the degrees of freedom is less than the total number of parameters, $df < q$.

In this thesis, we approximate the matrix $\mathscr{I}$ by the matrix $\mathbf{M}$ as defined in

Section 2.4.2. This implies that we use an approximation to the degrees of freedom.

3.4 Prediction

Once a multi-state model is fitted using a parametric and semi-parametric hazard model, estimated model parameters can be used for prediction. Typically this concerns computing transition matrices as a function of the penalised maximum likelihood estimate. The covariance of a function of model parameters can be estimated by Monte Carlo simulation or by using the multivariate delta method. Because transition probabilities are restricted to $[0,1]$, we have to calculate the standard error of e.g., $\log(p/(1-p))$ or $\log(-\log(p))$, and back transform (Titman, 2011). This thesis focuses on the simulation method.

Let $\widehat{\mathbf{V}}_{\boldsymbol{\theta}}$ denote the estimated covariance matrix of the penalised maximum likelihood estimate, $\widehat{\boldsymbol{\theta}}$, defined as $\mathbf{M}_p(\widehat{\boldsymbol{\theta}})^{-1}$ in (3.22). Notice that $\widehat{\mathbf{V}}_{\boldsymbol{\theta}}$ takes into account the choice of the smoothing parameters, $\widehat{\boldsymbol{\lambda}}$. Of interest is the estimation of $\mathbf{P}(t_1,t_2)$ for arbitrary t_1 and $t_2 > t_1$. In the case of a time-dependent model, let the grid for the piecewise-constant approximation be defined by $u_{j+1} = u_j + h$ for $j = 1,...,M$ such that $u_1 = t_1$ and $u_M = t_2$. Given this grid, matrix $\mathbf{P}(t_1,t_2)$ is estimated by $\mathbf{P}(u_1,u_2) \times \cdots \times \mathbf{P}(u_{M-1},u_M)$.

For Monte Carlo simulation, parameter vectors $\boldsymbol{\theta}^{(b)}$ are drawn from $N(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}_{\boldsymbol{\theta}})$, for $b = 1,...,B$, and for each sampled $\boldsymbol{\theta}^{(b)}$, $\mathbf{P}(t_1,t_2)$ is calculated. Summary statistics such as mean and covariance can be derived easily from the B realisations of $\mathbf{P}(t_1,t_2)$.

3.5 Application to the English longitudinal study of ageing data

The method is illustrated with an application to the ELSA data. As described in Section 1.6.2, the ELSA data are a random sample of size $N = 1000$ individuals. Of these 1000 individuals, 205 died during the follow-up with age at death available. The sample has 544 women and 456 men. Highest educational qualification is dichotomised for the current analysis according to years of formal education: fewer

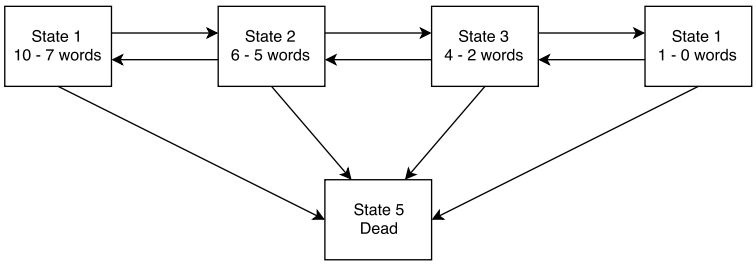

Figure 3.4: Five-state model for longitudinal data in ELSA on number of words remembered in a recall

Table 3.1: 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

than ten versus ten or more.

This application focuses on the number of words remembered in a delayed recall from a list of ten. A five-state model is defined for the number of words remembered and death, see Figure 3.4. The four living states are defined as: State 1, 2, 3, and 4, for the number of words $\{7,8,9,10\}$, $\{6,5\}$ $\{4,3,2\}$, and $\{1,0\}$, respectively. The state 5 is defined as the dead state. The statistical modelling in this section aims to explore the effect of age and gender on cognitive change over time when controlling for education. The interval-censored multi-state process is summarised by the frequencies in Table 3.1. Note that the sum of the transitions into the dead state is equal to the number of deaths in the sample, i.e., 205. Table 3.1 also shows that the process is mainly progressive in the sense that the main trend over time is towards the higher states.

In what follows, model estimation is undertaken by using the scoring algorithm. Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)^\top$ be the vector with model parameters, where q depends on the chosen model. The convergence criterion for the algorithm is to stop at iteration $v+1$ when $\sum_{k=1}^{q} |\theta_k^{(v)} - \theta_k^{(v+1)}| < 10^{-6}$. Because time of death is known, rather

than being interval censored, the likelihood contribution of individuals observed in state $r < 5$ at time t and dead at time $t^* > t$ are given as in (2.19) with $D = 5$.

Model selection is bottom-up starting with the time-homogeneous exponential hazard model given by

$$q_{rs}(t) = \exp\left(\beta_{rs.0}\right), \tag{3.25}$$

for the transitions $r \to s$ depicted in Figure 1.10. This intercept-only model with 10 parameters has AIC = 8109.5. Convergence of the scoring algorithm was reached after 14 iterations, using starting values $\beta_{rs.0} = -3$ for all the parameters.

As described in Section 2.3, time-dependent models are estimated by using a piecewise-constant approximation to the hazards. For the ELSA data, the mean length of follow-up times is 2.178 years with standard deviation of 0.855 and the median is 2 years. Assuming that change in transition intensities in relation to the frequency of observation can be assessed in intervals of approximately 2 years, we use the data to define the grid for the piecewise-constant approximation rather than imposing a fixed grid to calculate the likelihood contributions. For the process at hand, age is the most suitable time scale. Age in the ELSA data is transformed by subtracting 49 years. This results in 1 being the minimal age in the sample.

The first extension is a Gompertz models given by

$$q_{rs}(t) = \exp\left(\beta_{rs.0} + \xi_{rs}t\right), \tag{3.26}$$

where the effect of time is allowed to be different for all transitions. This model has 20 parameters and AIC = 7784.2.

Even though the sample size is not small, Table 3.1 shows that mortality information is limited because only about 20% of the individuals end up in the dead state during follow-up. A more parsimonious model Gompertz model is given by

$$q_{rs}(t) = \exp\left(\beta_{rs.0} + \xi_{rs}t\right), \tag{3.27}$$

where $\xi_{21} = \xi_{32} = \xi_{43} = \xi_B$ and $\xi_{15} = \xi_{25} = \xi_{35} = \xi_{45} = \xi_D$. That is, the effect of time is the same for all backwards transitions and for transitions into the dead state. This model has 15 parameters, and needs 16 scoring iterations when using starting values $\beta_{rs.0} = -3$ and $\xi_{rs} = 0$ for all the relevant r,s-combinations. The model has AIC = 7780.5. In what follows, model (3.25) is extended by adding parameters with parameter equality constraints.

Subsequently, covariate information is added for the transitions of interest, i.e., those transitions that represent a decline in cognitive function. For this, model (3.27) is extended to

$$q_{rs}(t) = \exp\left(\beta_{rs.0} + \xi_{rs}t + \beta_{rs.1}sex + \beta_{rs.2}education\right), \tag{3.28}$$

where *sex* is 0/1 for women/men, and *education* is 0/1 for fewer than ten years/ten years of more of education. For the transitions into the dead state, the constraints on the coefficients for *sex* are $\beta_{15.1} = \beta_{25.1} = \beta_{35.1} = \beta_{45.1}$, and for *education* are $\beta_{r5.2} = 0$ for $r = 1,2,3,4$. This model has 22 parameters, needs 16 iterations, and has AIC = 7680.3.

It is worthwhile to investigate alternative time-dependent models. First, in model (3.28), the Gompertz baseline models for the transitions into the dead state are replaced by Weibull models. Starting values for the transitions into the dead state are $\beta_{r5.0} = -10$, $\tau_{15} = \exp(0.5)$, and for the remaining parameters the values are as given above. This yields AIC = 7688.7 after 20 iterations.

Next, all baseline hazards definitions in model (3.28) are replaced by Weibull models, which results in AIC = 7729.5 after 28 iterations. Alternatively, model (3.28) is defined with Gompertz baseline models for the transitions into the dead state and Weibull models for progression through the living states. This yields AIC = 7719.7 after 25 iterations.

Semi-parametric models with *P*-splines can be used to model non-linear functional forms and to check shapes specified by parametric models. Because the focus of our investigation is the decline of cognitive function, which is mostly associated with individuals in states 3 and 4, we replace the Gompertz hazard for transition

Table 3.2: Comparison between models for the ELSA data with $N = 1000$, where -2LL stands for -2 times the (penalised) loglikelihood function evaluated at its maximum. The variable t denotes age transformed by subtracting 49 years

Model	Baseline hazards	#Parameters	-2LL	AIC
Intercept-only	Exponential	10	8089.5	8109.5
t	Gompertz no constraints	20	7744.2	7784.2
t	Gompertz	15	7750.5	7780.5
t, *sex*, *education*	Gompertz	22	7636.3	7680.3
t, *sex*, *education*	Gompertz for living and Weibull for death	22	7644.7	7688.7
t, *sex*, *education*	Weibull	22	7685.5	7729.5
t, *sex*, *education*	Weibull for living and Gompertz for death	22	7675.7	7719.7
t, *sex*, *education*	P-splines I for $2 \to 3$ and $3 \to 4$	38	7626.0	7678.2
t, *sex*, *education*	P-splines II for $3 \to 4$	30	7630.9	7678.2

$2 \to 3$ and $3 \to 4$ in model (3.28) by

$$\begin{aligned} q_{23}(t) &= \exp\left(\sum_{k=1}^{K} \alpha_{23.k} B_k(t) + \beta_{23.1} sex + \beta_{23.2} education\right) \\ q_{34}(t) &= \exp\left(\sum_{k=1}^{K} \alpha_{34.k} B_k(t) + \beta_{34.1} sex + \beta_{34.2} education\right). \end{aligned} \tag{3.29}$$

For this model, the number of P-splines bases for both hazard functions is $K = 10$ and the vector of smoothing parameters is $\lambda = (\lambda_1, \lambda_2)$. The initial grid is given by all pairs of combinations of $\log_{10}\lambda_1 = (-3,-2,-1,0,1,2,3)$ and $\log_{10}\lambda_2 = (-3,-2,-1,0,1,2,3)$. A possible graphical representation of the AIC results is to plot its values when one smoothing parameter is fixed. Figure 3.5(c) illustrates the resulting AIC for different values of λ_2 with fixed $\lambda_1 = 10^{-3}$. The value which minimises the AIC is $\lambda_2 = 10$. It happens for all values of λ_1. The search for the optimal values of λ_1 is less straightforward as $\lambda_1 \to \infty$. Figure 3.5(a) shows the AIC for several values of λ_1 with fixed $\lambda_2 = 10$. The AIC decreases quickly for small values of λ_2; however, it gets approximately constant for large values. This result indicates that the functional form of the hazard for transitions $2 \to 3$ is log-linear. Because both AIC and parameter estimates do not change much for sufficiently large values of λ_1, it is possible to set $\lambda_1 = 10^7$. In this case, the best

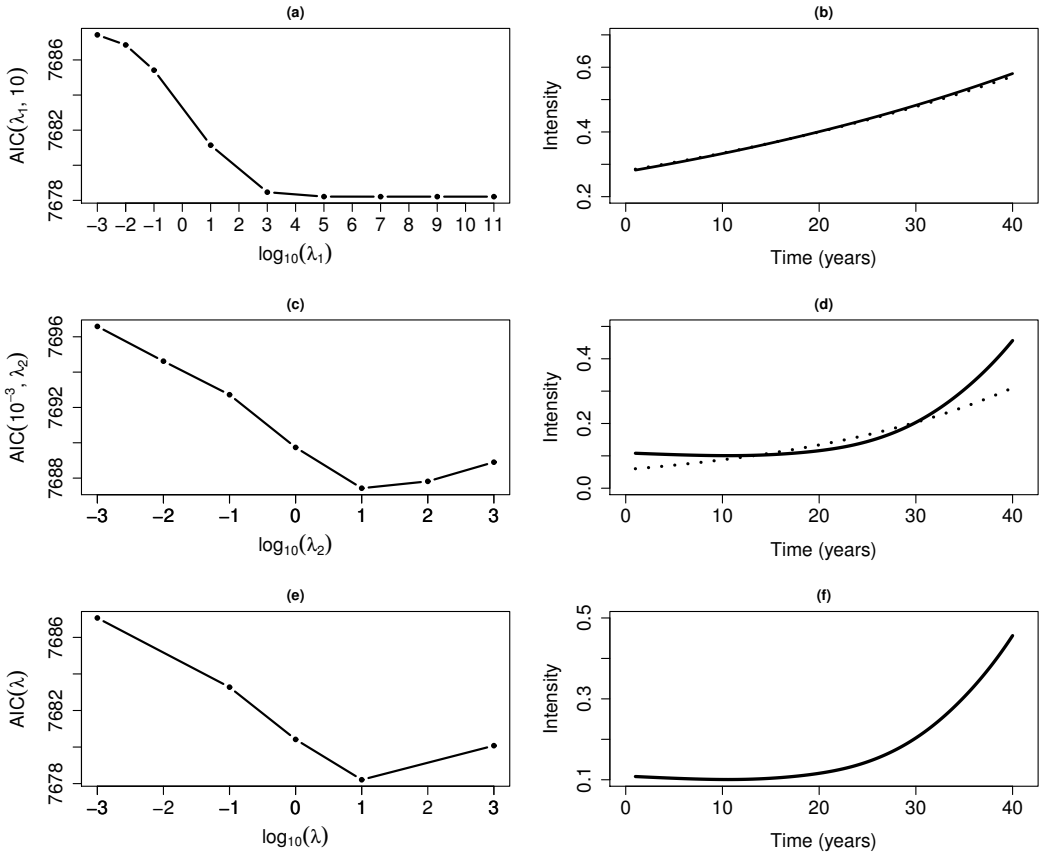

Figure 3.5: AIC results and fitted hazard transitions for men with ten or more year of education. In (a) and (c), the AIC results for fixed $\lambda_2 = 10$ and fixed $\lambda_1 = 10^{-3}$, respectively. In (b) and (d), the estimated hazards for $2 \to 3$ and $3 \to 4$, respectively. Solid line for *P*-splines I and dotted line for Gompertz. In (e) and (f), the AIC results for model *P*-splines II and fitted hazard for $3 \to 4$, respectively. Time denotes age transformed by subtracting 49 years

model (*P*-splines I) according to the AIC is obtained with smoothing parameter $\widehat{\lambda}^{\top} = (10^7, 10)$. This model has 30 parameters and 26.1 degrees of freedom.

The fitted hazards for transition $2 \to 3$ for the Gompertz (3.28) and *P*-splines I (3.29), for men with ten or more years of education are illustrated in Figure 3.5(b). The functional forms of both models are very similar for this transition; however, the functional forms for transition $3 \to 4$ are quite different, as indicated in Figure 3.5(d). Model (3.29) has AIC = 7678.2 indicating that it performs slightly better than the Gompertz model with AIC = 7680.3. Therefore, for prediction purposes the Gompertz models may be preferable for prediction beyond the time range in the data. The fitted hazards in Figures 3.5(b) and (d) show that an increase of age is

Table 3.3: Results for sex, education and time for the five-state *P*-splines II model for the ELSA data. Estimated standard errors in parentheses. The variable t denotes age transformed by subtracting 49 years

sex		*education*		t	
$\beta_{12.1}$	0.552 (0.138)	$\beta_{12.2}$	-0.281 (0.146)	ξ_{12}	0.030 (0.010)
$\beta_{23.1}$	0.178 (0.101)	$\beta_{23.2}$	-0.836 (0.103)	ξ_B	-0.031 (0.006)
$\beta_{34.1}$	0.141 (0.145)	$\beta_{34.2}$	-0.445 (0.160)	ξ_D	0.042 (0.009)
β_D	0.477 (0.151)				

associated with higher risk of moving from state 2 to state 3 and from state 3 to state 4.

The functional form of hazard for transition $2 \rightarrow 3$ in model (3.29) indicates that a Gompertz specification can be reasonable for this transition. Therefore, in model (3.28), only the hazard for transition $3 \rightarrow 4$ is specified with *P*-splines:

$$q_{34}(t) = \exp\left(\sum_{k=1}^{K} \alpha_{34.k} B_k(t) + \beta_{34.1} sex + \beta_{34.2} education\right). \quad (3.30)$$

The number of *P*-splines bases is $K = 10$ and the grid search is made on the values $\log_{10}\lambda = (-3,-1,0,1,3)$. The resulting AIC values are illustrated in Figure 3.5(e). The minimum AIC with value 7678.2 is obtained at $\lambda = 10$. That is the same AIC value as for model (3.29); however, the degrees of freedom is slightly smaller $df = 23.65$. As model (3.30) (*P*-splines II) is easier to estimate if compared to model (3.29), it is considered the best model among all illustrated in this thesis.

Table 3.2 summarises the comparison of the investigated models. Figure 3.5(f) illustrates the fitted hazard for transition $3 \rightarrow 4$ in model (3.30) for men with ten or more years of education. As expected, there is an increase in the risk of progression to a decline of cognitive function over the years.

Table 3.3 shows the estimates for the covariate effect parameters in the *P*-splines II model (3.30). Most of the point estimates are as expected. For example, the effect of getting older is associated with decline of cognitive function, i.e., $\widehat{\xi}_{12} > 0$, and with a decreasing hazard of remembering more words, i.e., $\widehat{\xi}_B < 0$. For transitions $1 \rightarrow 2$, $2 \rightarrow 3$, and $3 \rightarrow 4$ more years of education is associated with a

lower risk of moving. The effect of being a male patient is associated with higher risks of going into the dead state, $\widehat{\beta}_D = 0.477$. This correspond to hazard ratio of $\exp(\widehat{\beta}_D) = 1.611$, which represents a 61% higher hazard of death.

Model validation is hampered by the interval censoring of the transitions between the living states. But given that death times are available, it make sense to compare survival as estimated by the model with Kaplan-Meier curves (Gentleman et al., 1994). Of course, this will only check part of the fitted model. Figure 3.6 depicts baseline-specific survival as estimated by the model and as described by the Kaplan-Meier curves. Individual survival curves (in grey) are shifted to the *years since baseline* so that we can compare them and their mean to the Kaplan-Meier curves. This is necessary because individuals have different ages at baseline. The shape of the Kaplan-Meier curves is due to fact that time of death is rounded to the nearest integer. Even though time of death is not known precisely, we assume that time of transition into the dead state is known exactly. For survival given baseline state 3, there is some discrepancy between model-based mean survival and the Kaplan-Meier curve, but overall the fit is reasonably good. Although this is not a proper goodness-of-fit test, the comparison shows that the model is able to capture the attrition due to death during the follow-up.

3.5.1 Predicting cognitive function

Although parameters for the transition intensities help to understand the estimated model, interpretation is more straightforward when transition probabilities are considered. Firstly, consider a short time interval for which we assume that the intensities are constant. For men aged 60 with ten or more years of education, the two-year transition probabilities are estimated at

$$\widehat{\mathbf{P}}(t_1 = 11, t_2 = 13) = \begin{pmatrix} 0.330 & 0.488 & 0.154 & 0.010 & 0.018 \\ 0.171 & 0.531 & 0.253 & 0.023 & 0.022 \\ 0.083 & 0.391 & 0.429 & 0.066 & 0.031 \\ 0.034 & 0.219 & 0.410 & 0.291 & 0.046 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.31}$$

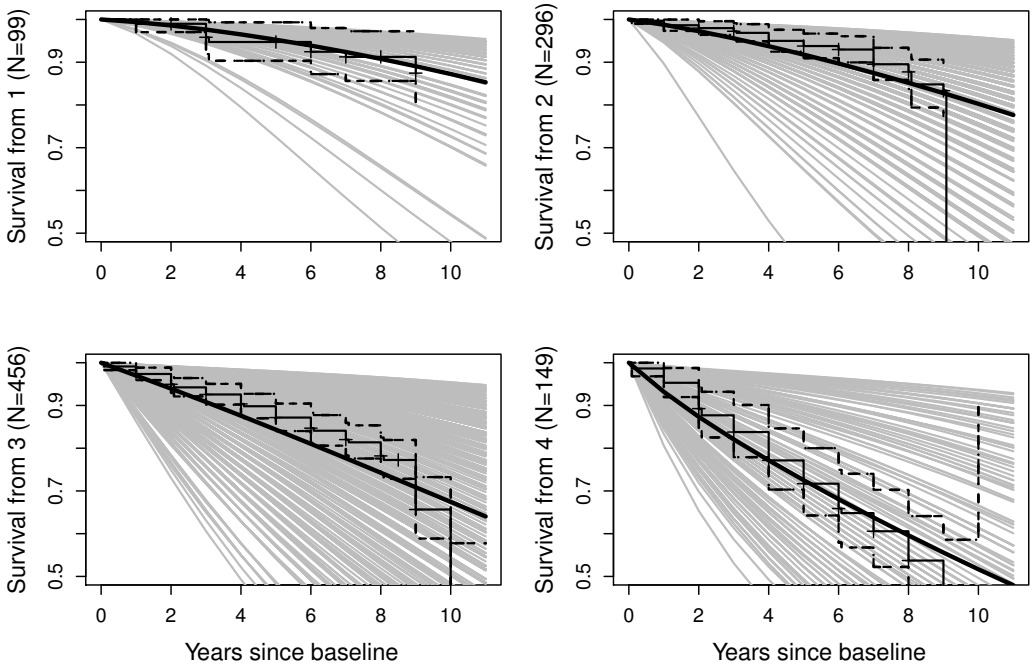

Figure 3.6: Comparison of model-based survival from states 1, 2, 3, and 4 with Kaplan-Meier curves. Model-based survival: grey lines for individuals, smooth black line for the mean of the individual survival curves. Kaplan-Meier in black lines with 95% confidence bands. Frequencies for baseline state along vertical axes

where t denotes age transformed by subtracting 49 years. The diagonal entries in this matrix dominate. But there are some large off-diagonal entries as well. For example, if a man aged 60 is in state 3, then he has a 39% chance of being in state 2 two years later. This high chance is an illustration of the noise of the process under investigation: it is quite likely that a 60 year old man moves between states 2 and 3 within the next two years.

Next we illustrate the estimation of standard errors and 95% confidence intervals for transition probabilities. Using simulation with $B = 1000$, the estimated standard errors of matrix (4.22) is

$$\begin{pmatrix} 0.038 & 0.029 & 0.016 & 0.004 & 0.011 \\ 0.013 & 0.021 & 0.019 & 0.009 & 0.006 \\ 0.007 & 0.019 & 0.024 & 0.024 & 0.006 \\ 0.004 & 0.019 & 0.024 & 0.037 & 0.012 \end{pmatrix}.$$

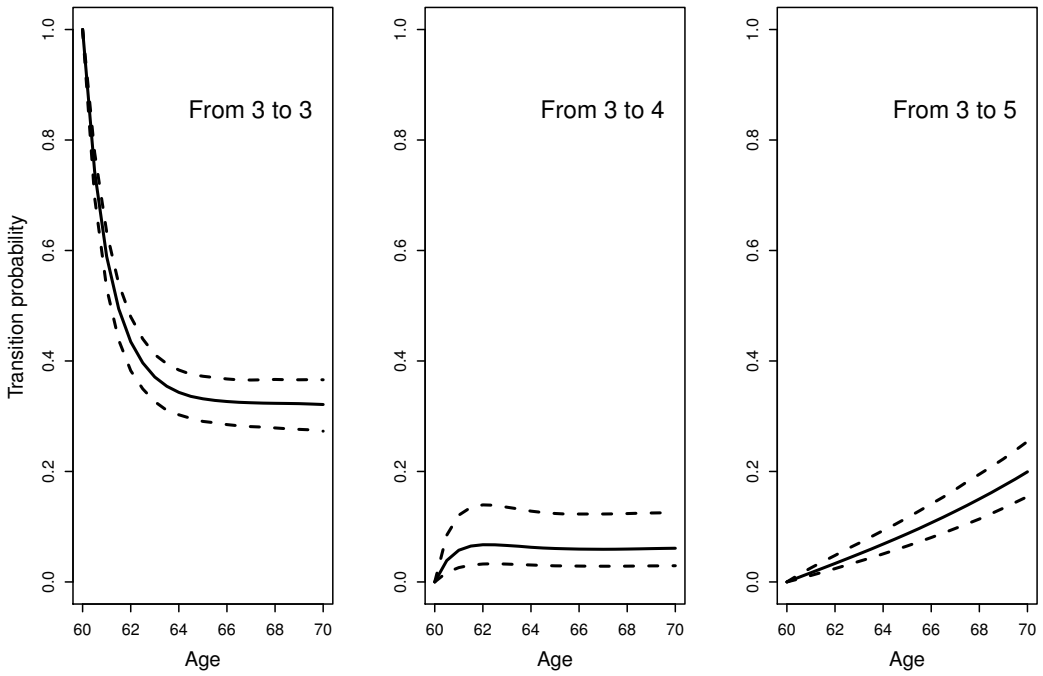

Figure 3.7: For the *P*-splines II model, estimated ten-year transition probabilities for men aged 60 with ten or more years of education, and in state 3 at baseline. Solid line for transition probabilities (with $B = 1000$) and dashed lines for 95% confidence bands

The 95% confidence intervals for the first row are given by

$$(0.263, 0.408),\ (0.425, 0.540),\ (0.120, 0.185),\ (0.004, 0.021), \text{ and } (0.011, 0.049).$$

Next, ten-year transition probabilities are estimated for men aged 60 with ten or more years of education. The grid is defined by $h = 1/2$ years. The estimation is shown in Figure 3.7.

Figure 3.7 concurs with expectations. For example, given the progressive trend of the process, it is to be expected that a probability of being in state 3 decreases over time, as moving to states 4, and 5 becomes more likely due to increased age.

3.6 Discussion

Specification and estimation of continuous-time multi-state models are presented and shown to be a flexible framework for statistical modelling of time-dependent processes. By defining transition-specific with parametric and semi-parametric

hazard models, a wide range of multi-state processes can be investigated. Penalised maximum likelihood estimation is undertaken by a scoring algorithm using a piecewise-constant approximation to time-dependent hazards. The Akaike information criterion is used to select the optimal value for the smoothing parameters.

The Markov processes formulation to semi-parametric multi-state models extends the method described in Joly and Commenges (1999) and Joly et al. (2002). This is an important methodology to medical statistics as backward transitions occur naturally in many applications (Abner et al., 2012; Marioni et al., 2012). Furthermore, using the piecewise-constant approximation is an alternative to the method introduced by Titman (2011) which handles the time-dependency by using numerical solutions to the non-linear differential equations defined directly by the time-dependency of the Markov process. As stated by Titman, computation using the non-linear differential equations can become prohibitively slow when adding continuous covariates. This is not a problem when using the piecewise-constant approximation and the scoring algorithm. To address the problem with continuous covariates, the method in Titman (2011) uses an approximation to the full likelihood. Therefore, It would be interesting to compare both methods to investigate the degree of approximation and relative computation speeds between the methods.

The piecewise-constant approximation can be determined by the observation times in the data. Alternatively, a fixed grid can be used that is imposed for all likelihood contributions. This method is computationally more extensive as it requires a greater number of eigenvalue decompositions to calculate transition probabilities for intervals defined by the grid. Those two methods to define a piecewise-constant approximation will differ depending on the study design and volatility of the process of interest (Van den Hout, 2017).

A semi-Markov assumption could be more reasonable for modelling the ELSA data; however, fitting semi-Markov models with interval-censored data is complicated given the number of living states and backward transitions (Commenges, 2002). Nonetheless, Figure 3.6 shows that the fit is reasonably good.

The scoring algorithm is implemented in R in such a way that it is easy to

vary transition-specific choices for parametric and semi-parametric shapes. An example of such a model is explored in the application, where *P*-splines are used for transitions $2 \rightarrow 3$ and $3 \rightarrow 4$, and Gompertz hazards are defined for the other transitions. The eigenvalue decomposition in the algorithm is computed with the function `eigen` in R, which uses the LAPACK routine (Anderson et al., 1999). *P*-spline bases are computed using the code in the appendix in Eilers and Marx (1996).

There is some overlap between the method presented in this chapter and the msm package (Jackson, 2011). The last is a platform to analyse time-homogeneous multi-state processes with interval-censored transition times. It is possible to fit some time-dependent models with msm, such as the Gompertz and splines models. However, this package cannot fit models with penalised splines nor with some commonly used parametric specifications such as the Weibull model.

If prediction of a time-dependent process beyond the time range in the data is of interest, hazard models with *P*-splines can be used to validate the parametric choices which underlie the prediction. This was illustrated in the application with the ELSA data in which age range is from 50 to 90 years. If risk factors are the main focus of the research, *P*-splines can be used to capture non-parametric shapes of time-dependency.

The choice of the type of spline is not essential. *P*-splines were used in this chapter, but any other spline function with a first-order derivative can be handled within the current framework. The same holds for parametric shapes other than the Gompertz and the Weibull. The specification and estimation of a continuous-time survival model is very general and does not pose restrictions on the number of states, scale of covariates, or number of transitions.

Chapter 4

Automatic smoothing for multi-state models

The method presented in the last chapter can become intractable for many applications, as computational time increases quickly with the number of hazards approached with splines. This chapter focus on developing an efficient and automatic method for estimating multi-state models with splines. Automatic smoothing methods are pivotal for practical modelling with splines. This chapter begins with a description of a commonly used method for multiple smoothing parameters estimation, and its limitation for multi-state models. Subsequently, we develop a penalised maximum likelihood method for automatic smoothing in multi-state models. The method is illustrated with applications to the CAV and OCTO data sets. A small simulation study is used to assess the performance of the method. A substantial part of the material in this chapter has been submitted to *arXiv* as a paper entitled *Penalised maximum likelihood estimation in multistate models for interval-censored data* (Machado et al., 2018).

4.1 Background for automatic smoothing

Chapter 3 shows that it is straightforward to specify parametric and semi-parametric hazard in multi-state models. For each transition-specific hazards approached with splines, a smoothing parameter is employed for estimation. Hence, the problem of estimating multi-state models with splines involves estimating multiple smoothing

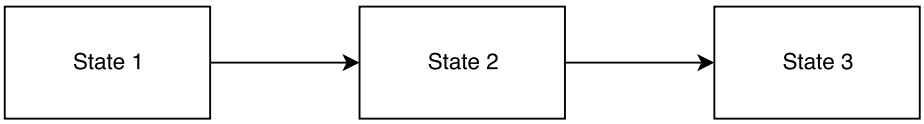

Figure 4.1: An unidirectional three-state model

parameters. Automatic smoothing parameters selection methods for generalised additive models (GAM) are well established (Ruppert et al., 2003; Wood, 2006). Those methods can be extended for different settings, e.g., semi-parametric copula regression models (Radice et al., 2016). In this section, we illustrate how the existing GAM framework could be applied to multi-state models, and discuss the limitations of such methods for multi-state modelling. For ease of presentation, the method is illustrated for the unidirectional three-state model in Figure 4.1. The details of most results in this section are presented in Section 4.2. The following presentation is based on Wood (2000) and Radice et al. (2016).

Let n represent the number of follow-up times in the data. The linear predictors for transitions $1 \rightarrow 2$ and $2 \rightarrow 3$ are given by

$$\eta_{12.i} = \mathbf{x}_i^\top \boldsymbol{\beta}_{12} + \sum_{k=1}^{K} \alpha_{12.k} B_{12.k}(t_i) \tag{4.1}$$

$$\eta_{23.i} = \mathbf{x}_i^\top \boldsymbol{\beta}_{23} + \sum_{k=1}^{K} \alpha_{23.k} B_{23.k}(t_i) \tag{4.2}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is a vector of covariates, $\boldsymbol{\beta}_{12} = (\beta_{12.1}, \ldots, \beta_{12.p})^\top$ and $\boldsymbol{\beta}_{23} = (\beta_{23.1}, \ldots, \beta_{23.p})^\top$ are parameter vectors. The splines parameters can be written as $\boldsymbol{\alpha}_{12} = (\alpha_{12.1}, \ldots, \alpha_{12.K})^\top$ and $\boldsymbol{\alpha}_{23} = (\alpha_{23.1}, \ldots, \alpha_{23.K})^\top$, where K represents the number of splines basis functions. The linear predictors can further be written as

$$\eta_{12.i} = \mathbf{x}_i^\top \boldsymbol{\beta}_{12} + \mathbf{B}_{12}(t_i)^\top \boldsymbol{\alpha}_{12} \tag{4.3}$$

$$\eta_{23.i} = \mathbf{x}_i^\top \boldsymbol{\beta}_{23} + \mathbf{B}_{23}(t_i)^\top \boldsymbol{\alpha}_{23}, \tag{4.4}$$

where $\mathbf{B}_{12}(t_i) = (B_{12.1}(t_i), \ldots, B_{12.K}(t_i))^\top$ and $\mathbf{B}_{23}(t_i) = (B_{23.1}(t_i), \ldots, B_{23.K}(t_i))^\top$. After defining, $\mathbf{X}_{1i} = (\mathbf{x}_i^\top, \mathbf{B}_{12}(t_i)^\top)^\top$ and $\mathbf{X}_{2i} = (\mathbf{x}_i^\top, \mathbf{B}_{23}(t_i)^\top)^\top$, we have that $\eta_{1i} =$

$\mathbf{X}_{1i}^\top\boldsymbol{\theta}_1$ and $\eta_{2i} = \mathbf{X}_{2i}^\top\boldsymbol{\theta}_2$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_{12}^\top, \boldsymbol{\alpha}_{12}^\top)^\top$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_{23}^\top, \boldsymbol{\alpha}_{23}^\top)^\top$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_{12}^\top, \boldsymbol{\alpha}_{12}^\top, \boldsymbol{\beta}_{23}^\top, \boldsymbol{\alpha}_{23}^\top)^\top$ be the full set of parameters. The penalised log-likelihood is given by

$$\ell_p(\boldsymbol{\theta}) \;=\; \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\theta}, \tag{4.5}$$

where $\ell(\boldsymbol{\theta})$ is the log-likelihood function and $\mathbf{S}_{\boldsymbol{\lambda}} = \text{diag}(\mathbf{0}, \lambda_1\mathbf{S}_1, \mathbf{0}, \lambda_2\mathbf{S}_2)$ is the penalty matrix.

For the estimation of the model parameters, given a value for the vector of smoothing parameters, $\widehat{\boldsymbol{\lambda}}$, we aim to find an update for the parameter vector $\boldsymbol{\theta}^{[a]}$. Let us define $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$, where $\mathbf{X}_i = \text{diag}\left(\mathbf{X}_{1i}^\top, \mathbf{X}_{2i}^\top\right)$, the vector $\mathbf{d}^{[a]}$ as a vector with ith element given by

$$\mathbf{d}_i^{[a]} = \left(\partial\ell(\boldsymbol{\theta}^{[a]})_i/\partial\eta_{12.i}, \partial\ell(\boldsymbol{\theta}^{[a]})_i/\partial\eta_{21.i}\right), \tag{4.6}$$

and $\mathbf{W}^{[a]}$ a block diagonal matrix made up of 2×2 matrices $\mathbf{W}_i^{[a]}$ given by

$$\begin{pmatrix} \dfrac{\partial^2\ell(\boldsymbol{\theta}^{[a]})_i}{\partial\eta_{12.i}\partial\eta_{12.i}} & \dfrac{\partial^2\ell(\boldsymbol{\theta}^{[a]})_i}{\partial\eta_{12.i}\partial\eta_{23.i}} \\ \dfrac{\partial^2\ell(\boldsymbol{\theta}^{[a]})_i}{\partial\eta_{23.i}\partial\eta_{12.i}} & \dfrac{\partial^2\ell(\boldsymbol{\theta}^{[a]})_i}{\partial\eta_{23.i}\partial\eta_{23.i}} \end{pmatrix}. \tag{4.7}$$

We then have that the penalised gradient and penalised Hessian at $\boldsymbol{\theta}^{[a]}$ are respectively given by

$$\mathbf{g}_p^{[a]} \;=\; \mathbf{X}^\top\mathbf{d}^{[a]} - \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\theta}^{[a]} \tag{4.8}$$

$$\mathscr{H}_p^{[a]} \;=\; -\mathbf{X}^\top\mathbf{W}^{[a]}\mathbf{X} - \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}. \tag{4.9}$$

Applying a first-order Taylor expansion to $\mathbf{g}_p^{[a+1]}$ around $\boldsymbol{\theta}^{[a]}$, setting the resulting expression to zero, we find that a new fit $\boldsymbol{\theta}^{[a+1]}$ is given by

$$\boldsymbol{\theta}^{[a+1]} = \left(\mathbf{X}^\top\mathbf{W}^{[a]}\mathbf{X} + \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\right)^{-1}\mathbf{X}^\top\mathbf{W}^{[a]}\mathbf{z}^{[a]}, \tag{4.10}$$

where $\mathbf{z}^{[a]} = (\mathbf{W}^{[a]})^{-1}\mathbf{d}^{[a]} + \mathbf{X}\boldsymbol{\theta}^{[a]}$ is the pseudo-data (Wood, 2000). The derivation of this result is presented with more details in Section 4.2. The pseudo-data plays the role of the response vector in the GAM framework.

The smoothing parameter vector can be estimated by minimising the Un-Biased Risk Estimator given by

$$\mathscr{V}(\boldsymbol{\lambda}) = ||\mathbf{W}^{1/2}(\mathbf{z} - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z})||^2/\check{n} - 1 + 2tr(\mathbf{A}_{\boldsymbol{\lambda}})/\check{n}, \tag{4.11}$$

where $\mathbf{A}_{\boldsymbol{\lambda}} = \sqrt{\mathbf{W}}\mathbf{X}\left(\mathbf{X}^\top\mathbf{W}\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}}\right)^{-1}\mathbf{X}^\top\sqrt{\mathbf{W}}$ is the influence matrix, and $\check{n} = 2 \times n$ (Craven and Wahba, 1978). Equation (4.11) can be minimised using a method developed by Wood (2004), which is based on Newton's method and can evaluate the components of $\mathscr{V}(\boldsymbol{\lambda})$ and their first and second order derivatives.

In the derivation above, the $\mathbf{W}_i^{[a]}$ can be approximated by the $\mathbf{M}_i^{[a]}$ matrix defined as

$$\begin{pmatrix} \frac{\partial \ell(\boldsymbol{\theta}^{[a]})_i}{\partial \eta_{12.i}}\frac{\partial \ell(\boldsymbol{\theta}^{[a]})_i}{\partial \eta_{12.i}} & \frac{\partial \ell(\boldsymbol{\theta}^{[a]})_i}{\partial \eta_{12.i}}\frac{\partial \ell(\boldsymbol{\theta}^{[a]})_i}{\partial \eta_{23.i}} \\ \frac{\partial \ell(\boldsymbol{\theta}^{[a]})_i}{\partial \eta_{23.i}}\frac{\partial \ell(\boldsymbol{\theta}^{[a]})_i}{\partial \eta_{12.i}} & \frac{\partial \ell(\boldsymbol{\theta}^{[a]})_i}{\partial \eta_{23.i}}\frac{\partial \ell(\boldsymbol{\theta}^{[a]})_i}{\partial \eta_{23.i}} \end{pmatrix}, \tag{4.12}$$

for $i = 1, \ldots, n$.

The above penalised maximum likelihood estimation for multi-state models is based on the pseudo-data $\mathbf{z} = (\mathbf{W})^{-1}\mathbf{d} + \mathbf{X}\boldsymbol{\theta}$. The main limitation of this method is that the weight matrix $\mathbf{W}$ must be positive definite. For highly flexible models, not all the n weight matrices contained in $\mathbf{W} = \text{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$ need to be positive definite. Also, the matrix $\mathbf{W}$ is as large as there are data and transitions in a multi-state process. Then computational time and stability become serious issues for most applications.

Marra et al. (2017) propose a smoothing parameter estimation based on a parametrisation of $\mathbf{z}$ that uses $\mathscr{H}$ and $\mathbf{g}$ as a whole instead of the n components that make them up. For some applications, the Hessian $\mathscr{H}$ may not be positive definite but these would occur considerably less frequently than when working with the n weight matrices that make it up.

We next present a penalised maximum likelihood estimation for multi-state models with splines, that is based on the framework developed by Marra et al. (2017).

4.2 Penalised maximum likelihood estimation

This section presents an automatic and efficient method to estimate multi-state models with splines in the presence of interval censoring. We recall model specification and the piecewise-constant approximation to the time-dependency. The method presented in Marra et al. (2017) is then adapted for multi-state models with splines.

4.2.1 Model representation

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

$$q_{rs}(t) = q_{rs.0}(t) \exp\left(\boldsymbol{\beta}_{rs}^{\top}\mathbf{x}\right), \tag{4.13}$$

where $q_{rs.0}(t)$ is an unspecified baseline hazard function, $\mathbf{x} = (x_1, \ldots, x_p)^{\top}$ is a covariate vector and $\boldsymbol{\beta}_{rs} = (\beta_{rs.1}, \ldots, \beta_{rs.p})^{\top}$ is a vector of unknown parameters. The nonparametric specification of $q_{rs.0}(t)$ with splines is given by

$$q_{rs.0}(t) = \exp\left(\sum_{k=1}^{K_{rs}} \alpha_{rs.k} B_k(t)\right), \tag{4.14}$$

where K_{rs} is the number of spline base functions $B_k(t)$, and $\alpha_{rs.k} \in \mathbb{R}$ are regression coefficients.

4.2.2 Piecewise-constant hazards

Similarly to Chapter 3, time-dependency of the hazards are taken into account by using a piecewise-constant approximation. If the follow-up times vary across individuals, 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 (2008a). For most applications, both methods lead to similar results and the method described in this Section is preferable as it is less computationally extensive (Van den Hout, 2017).

4.2.3 Penalised log-likelihood function

For each hazard, let the number of splines basis dimension be large enough to allow for flexible modelling. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$ be the full set of parameters and $\ell(\boldsymbol{\theta})$ be the logarithm of the likelihood function. The smoothness of the model is controlled by adding a smoothness penalty to the log-likelihood function. The penalised log-likelihood function is

$$\ell_p(\boldsymbol{\theta}) \;=\; \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\theta}, \tag{4.15}$$

where $\mathbf{S}_{\boldsymbol{\lambda}}$ is the penalty matrix. This is a block diagonal matrix with blocks $\lambda_{rs}\mathbf{S}_{rs}$ for penalising splines parameters of transition r to s, where $\mathbf{S}_{rs}$ is a matrix of known coefficients.

4.2.4 Parameter estimation

Given a piecewise-constant approximation to the time-dependency in the hazard model (4.13), a scoring algorithm can be used to maximise the penalised log-likelihood function (4.15), see Section 3.3.2. 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.

As discussed in Section 4.1, there are methods available for automatic smoothing parameters estimation within the penalised likelihood framework (Wood, 2006; 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.

In this section, we present the method for automatic smoothing developed by

Marra et al. (2017), which uses the gradient and the Hessian (or Fisher information matrix) as a whole instead of components that make them up. The method consists 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 $\mathscr{H}_p^{[a]} = \mathscr{H}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}}$ 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 $\mathscr{H}^{[a]} = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[a]}}$. A first-order Taylor expansion of $\mathbf{g}_p^{[a+1]}$ about the current fit $\boldsymbol{\theta}^{[a]}$ is given by

$$\mathbf{g}_p^{[a+1]} \approx \mathbf{g}_p^{[a]} + \mathscr{H}_p^{[a]}(\boldsymbol{\theta}^{[a+1]} - \boldsymbol{\theta}^{[a]}), \tag{4.16}$$

where $\mathbf{g}_p^{[a+1]} = \mathbf{g}^{[a]} - \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\theta}^{[a]}$ and $\mathscr{H}_p^{[a]} = \mathscr{H}^{[a]} - \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}$. Let us define $\mathscr{I}^{[a]} = -\mathscr{H}^{[a]}$. A new fit $\boldsymbol{\theta}^{[a+1]}$ is obtained by taking the right-hand side of equation (4.16) to be zero

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

Therefore, for fixed value of $\widehat{\boldsymbol{\lambda}}$ the new fit for the parameter estimator can be expressed as

$$\boldsymbol{\theta}^{[a+1]} = \left(\mathscr{I}^{[a]} + \mathbf{S}_{\widehat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\mathscr{I}^{[a]}}\mathbf{z}^{[a]}, \tag{4.17}$$

where $\mathbf{z}^{[a]} = \sqrt{\mathscr{I}^{[a]}}\boldsymbol{\theta}^{[a]} + \boldsymbol{\varepsilon}^{[a]}$ is a $q \times 1$ vector, with $\boldsymbol{\varepsilon}^{[a]} = \sqrt{\mathscr{I}^{[a]}}^{-1}\mathbf{g}^{[a]}$ also a $q \times 1$ vector. The quantities $\mathbf{z}$ is called the *pseudo-data*, which plays the role of the response vector for GAM.

This parametrisation of the model-parameter estimators allows for a well founded formulation of the smoothing parameter selection that is presented in Section 4.2.5 (Marra et al., 2017). As discussed in Kalbfleisch and Lawless (1985), calculating the second derivatives of the probability matrix can be difficult to calculate. We use the approximation to the Fisher information matrix that involves only the first order derivatives of the penalised log-likelihood function as in (2.23).

4.2.5 Smoothing parameters estimation

The penalised maximum likelihood approach described in Section 4.2.4 can only estimate model parameters, $\boldsymbol{\theta}$, for fixed vector of smoothing parameters, $\boldsymbol{\lambda}$. In this Section, we described the automatic smoothing parameter selection presented in Marra et al. (2017).

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

$$\begin{aligned}
\mathbb{E}\left(||\boldsymbol{\mu}_\mathbf{z} - \widehat{\boldsymbol{\mu}_\mathbf{z}}||^2\right) &= \mathbb{E}\left(||(\mathbf{z} - \boldsymbol{\varepsilon}) - \mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{z}||^2\right) \\
&= \mathbb{E}\left(||(\mathbf{z} - \mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{z}) - \boldsymbol{\varepsilon}||^2\right) \\
&= \mathbb{E}\left(||\mathbf{z} - \mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{z}||^2 + ||\boldsymbol{\varepsilon}||^2 - 2\boldsymbol{\varepsilon}^\top(\mathbf{z} - \mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{z})\right) \\
&= \mathbb{E}\left(||\mathbf{z} - \mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{z}||^2 + \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - 2\boldsymbol{\varepsilon}^\top(\boldsymbol{\mu}_\mathbf{z} + \boldsymbol{\varepsilon}) + 2\boldsymbol{\varepsilon}^\top\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}(\boldsymbol{\mu}_\mathbf{z} + \boldsymbol{\varepsilon})\right) \\
&= \mathbb{E}\left(||\mathbf{z} - \mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{z}||^2 - \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - 2\boldsymbol{\varepsilon}^\top\boldsymbol{\mu}_\mathbf{z} + 2\boldsymbol{\varepsilon}^\top\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\mu}_\mathbf{z} + 2\boldsymbol{\varepsilon}^\top\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\varepsilon}\right).
\end{aligned}$$

Notice that $\mathbb{E}(\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}) = \mathbb{E}(\sum\varepsilon_i^2) = q$, $\mathbb{E}(\boldsymbol{\varepsilon}^\top\boldsymbol{\mu}_\mathbf{z}) = \mathbb{E}(\boldsymbol{\varepsilon}^\top)\boldsymbol{\mu}_\mathbf{z} = 0$ and $\mathbb{E}(\boldsymbol{\varepsilon}^\top\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\mu}_\mathbf{z}) = \mathbb{E}(\boldsymbol{\varepsilon}^\top)\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\mu}_\mathbf{z} = 0$. Using that $\boldsymbol{\varepsilon}^\top\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\varepsilon}$ is a scalar, and that a scalar is its own trace, we obtain that

$$\mathbb{E}(\mathrm{tr}(\boldsymbol{\varepsilon}^\top\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\varepsilon})) = \mathbb{E}(\mathrm{tr}(\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)) = \mathrm{tr}(\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)) = \mathrm{tr}(\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{I}) = \mathrm{tr}(\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}).$$

These calculations can also be found in Wood (2006). Hence, it follows that

$$\mathbb{E}\left(||\boldsymbol{\mu}_{\mathbf{z}} - \widehat{\boldsymbol{\mu}_{\mathbf{z}}}||^2\right) = \mathbb{E}\left(||\mathbf{z} - \mathbf{A}_{\widehat{\boldsymbol{\lambda}}}\mathbf{z}||^2\right) - q + 2\text{tr}(\mathbf{A}_{\widehat{\boldsymbol{\lambda}}}) \tag{4.18}$$

where q is the number of parameters. The derivation of (4.18) is presented with a minor mistake in Marra et al. (2017), where the constant q does not represent the total number of parameters.

Calculating the expectation on the right hand side of (4.18) is not straightforward. In practice, $\boldsymbol{\lambda}$ is estimated by minimising the Un-Biased Risk Estimator (UBRE; Craven and Wahba (1978))

$$\mathscr{V}(\boldsymbol{\lambda}) = ||\mathbf{z} - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}||^2 - q + 2tr(\mathbf{A}_{\boldsymbol{\lambda}}). \tag{4.19}$$

Equation (4.19) can be minimised using the automatic smoothing parameter selection method developed by (Wood, 2004) or by using a general-purpose optimiser.

4.2.6 Summary of the algorithm

The methods described in Sections 4.2.4 and 4.2.5 are iterated until the parameter estimator satisfies $\max_{1\leq k\leq q}|\theta^{[a+1]} - \theta^{[a]}| < \delta$ for a suitable small positive value δ (Radice et al., 2016). The two steps of the algorithm are as follows:

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

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

Step 2: Given the estimate $\boldsymbol{\theta}^{[a+1]}$, find an estimate of $\boldsymbol{\lambda}$ using equation (4.19):

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

4.2.7 Confidence intervals

The distribution of the penalised maximum likelihood estimator can be used to construct confidence intervals for the estimate $\widehat{\boldsymbol{\theta}}$ and functions of them, such as the

hazards and probability matrix (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 thesis, we approximate the covariance matrix $\mathbf{V}_{\boldsymbol{\theta}}$ by the inverse of the matrix $\mathbf{M}_p$ as defined in (3.22). The standard errors reflect the choice of theta, but do not account for the way theta was chosen, i.e., the same standard error would occur if theta were taken to be fixed and known from the outset.

4.3 Simulation study

We perform a small simulation study to analyse the performance of the method presented in Section 4.2 for estimating multi-state models with splines. The simulation is described for an illness-death model without recovery with a log-normal distribution with parameters $\mu = 1.25$ and $\sigma = 1$ for transition 1 to 2, an exponential distribution with rate $\exp(-2.5)$ for transition 1 to 3, and a Gompertz distribution with rate $\exp(-2.5)$ and shape 0.1 for transition 2 to 3.

Let $T_{rs} = T_{rs|u}$ represent the time to the event 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 if $T = T_{13}$ then the next state is 3. If the state is 2, then the time of the next state is T_{23}. The event times T_{12} and T_{13} are simulated using the functions rgengamma() and rgompertz(), respectively, in the package flexsurv (Jackson, 2016). The transition times T_{23} can be simulated by sampling from uniform distribution and using the inversion method as described in Section 1.5.

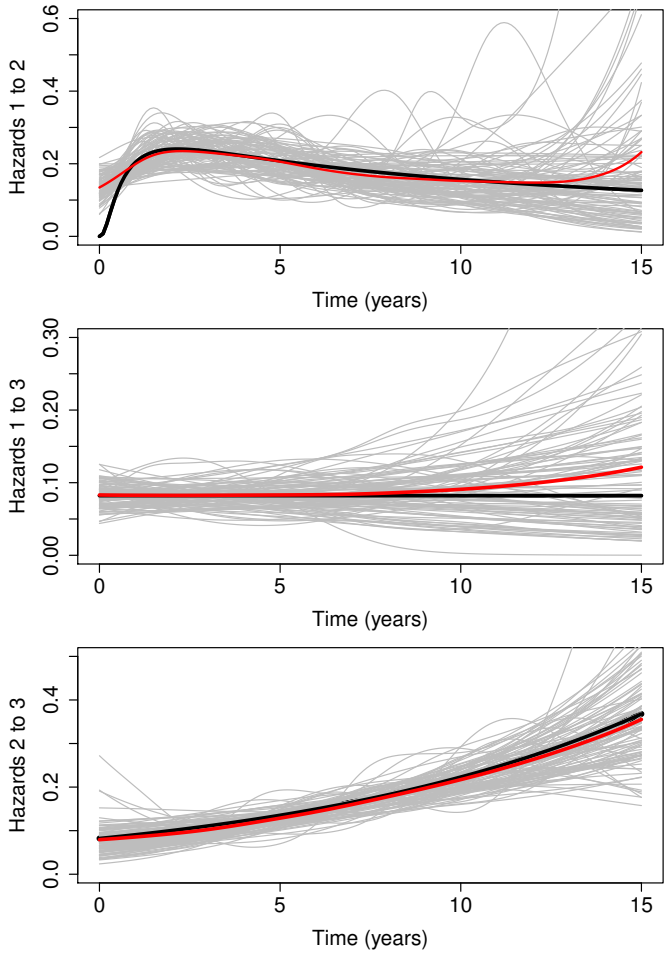

Figure 4.2: Simulation study: true (black lines), estimated (grey lines) and mean estimated (red lines) hazards for the illness-death model for 100 replications

We perform $R = 100$ replications. The sample size is $N = 200$ individuals. The time scale is years since baseline, i.e., time since the beginning of the study. The follow-up times are yearly and the length of the study is 15 years. This leads to interval-censored transition times for transitions 1 to 2 and known time of transitions into the dead state.

The package mgcv in R (Wood, 2007) is used to set the design and penalty matrices. The number of knots for each hazard is $K = 10$, hence the model has a total of 30 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 multi-state model with splines is then estimated using the procedure described in Section 4.2.

Figure 4.2 illustrates the true (black lines), estimated (grey lines) and mean estimated (red lines) hazards for the illness-death model for 100 replications. Some estimated hazards seem to over- or under-estimated the true hazards; however, the mean estimated hazards are very close to the true hazard curves. The discrepancy between the hazards towards the end of the study is due to scarcity of data after 10 years. Also, the bias for hazard for transition 1 to 2 at early follow-up times is due to the lack of transitions from state 1 to state 2 in the first year of the study. Overall, the method satisfactorily estimates the nonlinear trend underlying the hazard for transition 1 to 2.

Table 4.1 presents the results of the simulation in terms of transition probabilities. For each defined time interval, it shows the transition probabilities for the true model, the mean estimated transition probabilities, the bias and the estimated standard errors (eSE) (Burton et al., 2006). The results show that the multi-state model with splines can estimate well transition probabilities across all time intervals.

The findings from the simulation results are twofold. First, they indicate that the proposed method is able to estimate nonlinear 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 transition probabilities. Of course, the small number of replications prevent us from being able to assess the coverage of confidence intervals; however, the simulation results shows the good performance of the method with relation to recovering the true model.

4.4 Applications

The methods presented in this chapter are illustrated with applications to the OCTO and CAV data sets. In what follows, model estimation is undertaken using the scoring algorithm summarised in Section 4.2.6. Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)^\top$ be the vector with model parameters, where q depends on the application. The convergence criterion for the algorithm is to stop at iteration $a+1$ when $\max_{1 \leq k \leq q} |\theta_k^{[a]} - \theta_k^{[a+1]}| < 10^{-6}$.

We use penalised cubic regression splines for the basis function. The knots are

Table 4.1: Simulation study to investigate the performance of the multi-state models with splines for modelling time-dependent processes. Mean, bias and estimated standard errors (eSE) for $R = 100$ replications. Absolute bias less than x is denoted by $\lceil x \rceil$

Transition probabilities	True	Mean	Bias	eSE
$p_{11}(0,5)$	0.241	0.232	-0.090	0.03
$p_{12}(0,5)$	0.387	0.402	0.015	0.033
$p_{13}(0,5)$	0.372	0.366	-0.006	0.028
$p_{22}(0,5)$	0.589	0.611	0.022	0.057
$p_{23}(0,5)$	0.411	0.389	-0.022	0.057
$p_{11}(0,10)$	0.065	0.065	$\lceil 0.001 \rceil$	0.015
$p_{12}(0,10)$	0.232	0.239	0.008	0.023
$p_{13}(0,10)$	0.703	0.695	-0.008	0.028
$p_{22}(0,10)$	0.246	0.263	0.018	0.037
$p_{23}(0,10)$	0.754	0.737	-0.018	0.037
$p_{11}(5,10)$	0.269	0.281	-0.012	0.052
$p_{12}(5,10)$	0.291	0.284	0.007	0.037
$p_{13}(5,10)$	0.440	0.435	-0.005	0.052
$p_{22}(5,10)$	0.417	0.430	0.013	0.037
$p_{23}(5,10)$	0.583	0.570	-0.013	0.037

placed considering the percentiles of the observation times. This means that more knots are placed where data are plentiful and fewer knots where data are scarce. 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 might not be possible for some applications as it requires the knots to be equally spaced. In this case, some knots can be placed where there is no data. The design and penalty matrices are set up using the package mgcv in R. The smoothing parameters are estimated using the general-purpose optimiser optim in R.

4.4.1 Origins of variance in the oldest-old data

We showed in Section 2.8.1 that a Gompertz hazards specification seems to be reasonable for the OCTO data. In this section, these data are analysed with a multi-model with splines. The multi-state process is illustrated in Figure 4.3. We aim to check the suitability of the parametric hazards specification in Section 2.8.1, as

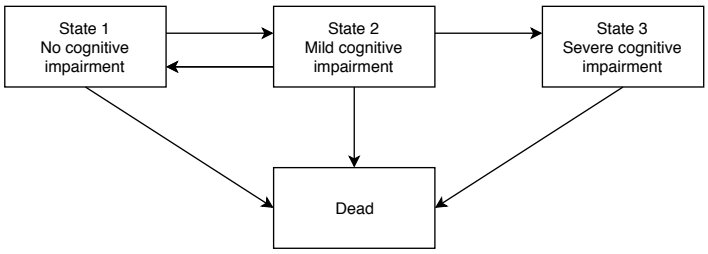

Figure 4.3: Four-state model for longitudinal data in OCTO

well as verify whether the modelling can be further improved with nonparametric hazards.

Recall that for the OCTO data, the mean length of follow-up times is 1.994 years with standard deviation of 1.098 and median 1.986. Assuming that change in transition intensities in relation to the frequency of observation can be assessed in intervals of approximately 2 years, we can use the data to define the grid for the piecewise-constant approximation.

Let t represent age minus 80. Because time of death is known, rather than being interval censored, the likelihood contribution of individuals observed in state $r < 4$ at time t and dead at time $t^* > t$ are given by $\sum_{s=1}^{3} P(Y(t^*) = s|Y(t) = r)\, q_{s4}(t^*)$.

Similarly to the application in Section 2.8.1, the proportional hazard model with splines and dependence on the covariate *sex* is given by

$$q_{rs}(t) = \exp\left(\sum_{k=1}^{10} \alpha_{rs.k} B_k(t) + \beta_{rs} sex\right), \tag{4.20}$$

where $(r,s) \in \{(1,2),(1,4),(2,1),(2,3),(2,4),(3,4)\}$, $B_k(t)$ are cubic regression splines, and *sex* is $0/1$ for men/women. For the transition between the living states, the constraints on the coefficients for *sex* are $\beta_{12.1} = \beta_{23.1} \stackrel{d}{=} \beta_L$, except for transition from 2 to 1 with $\beta_{21.1} = 0$. For the transitions into the dead state, the constraints are $\beta_{14.1} = \beta_{24.1} = \beta_{34.1} \stackrel{d}{=} \beta_D$.

As indicated in (4.20), the hazards are modelled with 10 knots each, hence the total number of parameters is 62. The resulting model has $AIC = 5340.946$, $-2\ell(\boldsymbol{\theta}) = 5306.396$, and $df = 17.275$. The comparison with the parametric model in (2.27), which has $AIC = 5342.837$, $-2\ell(\boldsymbol{\theta}) = 5314.837$, and 14 parameters, indi-

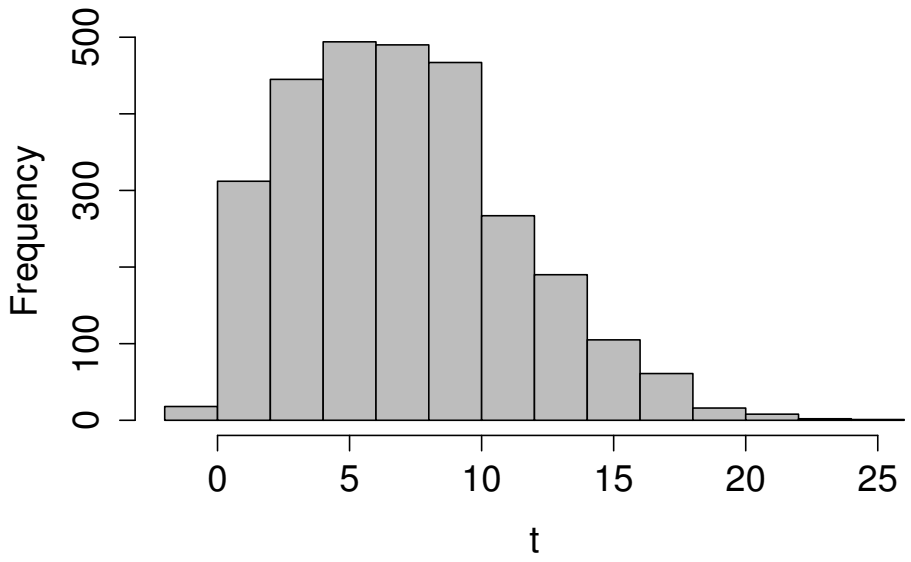

Figure 4.4: Histogram of age transformed by minus 80 in the OCTO data

cates that the nonparametric hazards specification in (4.20) leads to a more flexible model, which slightly improves model fitting.

The estimated smooth hazards for women (solid lines) and 95% confidence intervals (dashed lines) are presented in Figure 4.5. For transition from state 1 to 2, the hazard is increasing and has a nonlinear shape. For transition from state 2 to 1, the hazard slightly increases up to approximately 5 years, but decreases afterwards. The shapes of these hazards can be difficult to model with parametric specifications. For the other transitions, the risks of moving across states are increasing throughout the length of the study, with fairly log-linear shapes that can be approached with parametric models. The confidence intervals are fairly wide after approximately 15 years. That is because data become scarce after 15 years and our smooth non-parametric hazard model is a strictly local approach. Figure 2.1 illustrates the histogram of age minus 80. The vector of smoothing parameters is estimated at $\widehat{\boldsymbol{\lambda}} = (327.57,\ 640.20,\ 216.54,\ 3935634.85,\ 500128.59,\ 2118.19)^{\top}$.

The covariate effects are estimated at $\widehat{\beta_L} = -0.323\ (0.100)$ and $\widehat{\beta_D} = -0.338\ (0.092)$, indicating that being a woman decrease the risks of moving across

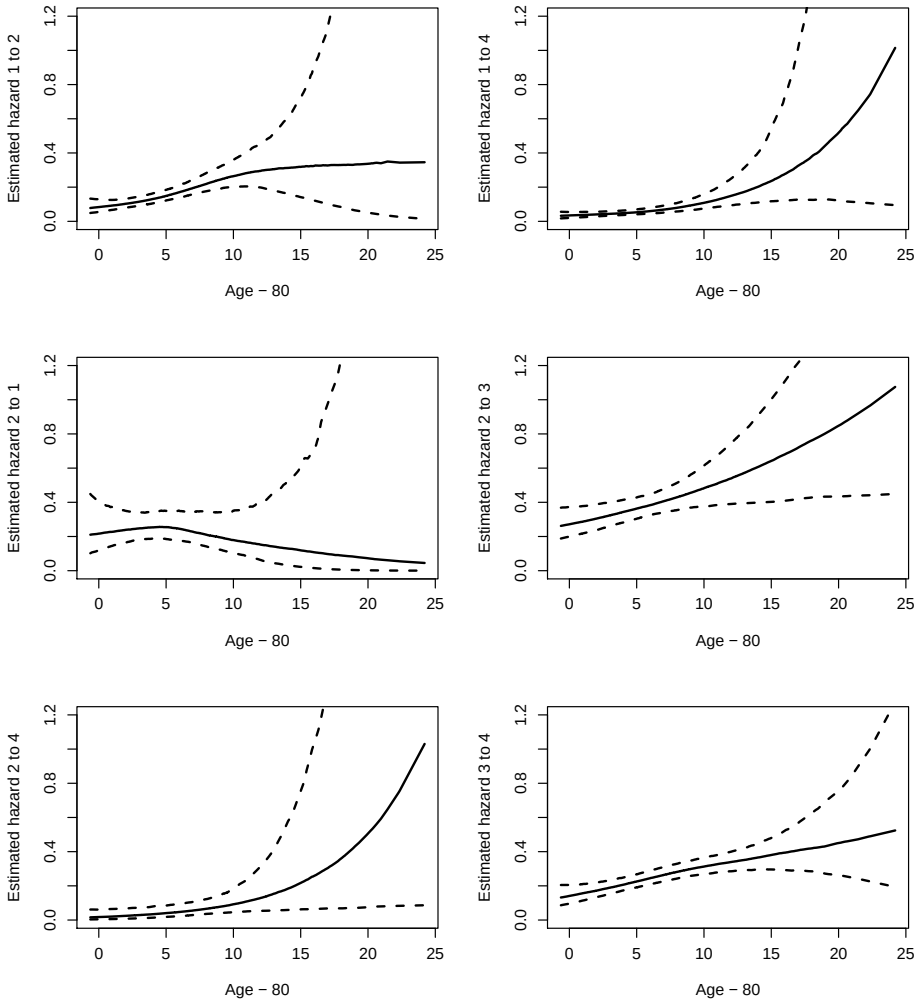

Figure 4.5: Estimated smooth hazards (solid lines) for women, with 95% confidence intervals (dashed lines)

cognitive states and moving into the dead state, respectively. Notice that for the parametric model in (2.27), the effect of the covariate *sex* for living states and into the dead state are estimated at -0.325 and -0.334, respectively. Therefore, the baseline hazard specification does not seem to influence the covariate estimates.

Figure 3.6 depicts baseline-specific survival as estimated by the model and as described by the Kaplan-Meier curves. Individual survival curves (in grey) are shifted to the *years since baseline* so that we can compare them and their mean to the Kaplan-Meier curve. This is necessary because individuals have different ages at baseline. For survival given baseline state 3, there is small discrepancy between model-based mean survival and the Kaplan-Meier curve, but overall the fit seems to

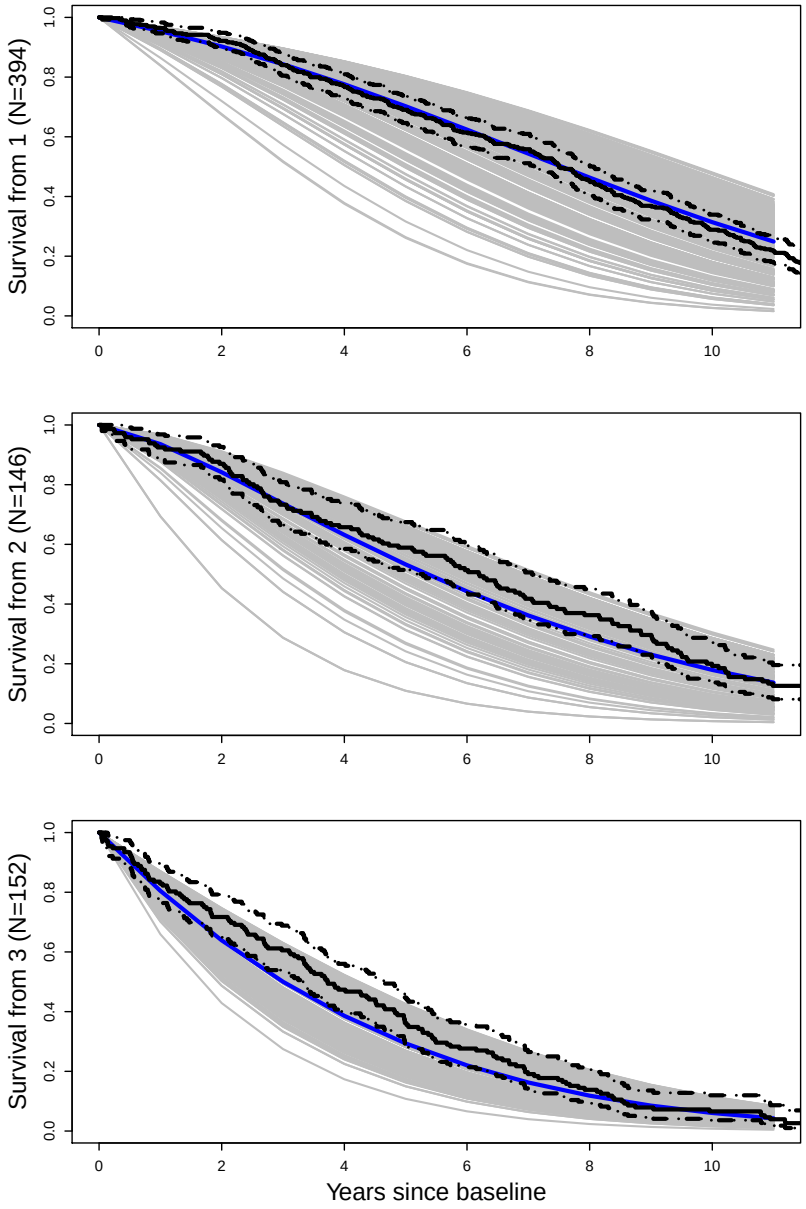

Figure 4.6: Comparison of model-based survival from states 1, 2, and 3 with Kaplan-Meier curves. Model-based survival: grey lines for individuals, smooth blue lines for the mean of the individual survival curves. Kaplan-Meier in black lines with 95% confidence intervals. Frequencies for baseline state along vertical axes

be good. Although this is not a proper goodness-of-fit test, the comparison shows that the model is able to capture the attrition due to death during the follow-up. Furthermore, the predictions for the nonparametric model (4.20) are better than the ones obtained for model (2.27), see Figure 2.3 .

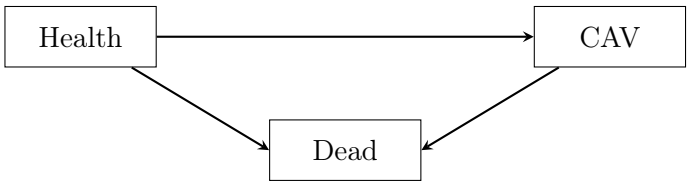

Figure 4.7: Illness-death without recovery model for disease progression after transplant

4.4.2 Cardiac allograft vasculopathy data

As illustrated in Section 2.8.2, Gompertz hazards specification for analysing the CAV data results in poor model fitting. In this section, we analyse these data with nonparametric hazards models. The multi-state process is illustrated in Figure 4.7. We aim to verify whether flexible model specifications with splines can improve model fitting.

As discussed in Section 2.8.2, the mean length of follow-up times is 1.622 years with standard deviation of 0.972 and median 1.258. Assuming that change in transition intensities in relation to the frequency of observation can be assessed in intervals of approximately 1 year, we can use the data to define the grid for the piecewise-constant approximation.

Let t represent time since baseline. Since 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 $\sum_{s=1}^{2} P(Y(t^*) = s|Y(t) = r)\, q_{s3}(t^*)$.

The proportional hazard model with splines is specified with dependence on donor age (*dage*) and primary diagnosis of ischaemic heart disease (*IHD*):

$$q_{rs}(t) = \exp\left(\sum_{k=1}^{7} \alpha_{rs.k} B_k(t) + \beta_1 dage + \beta_2 IHD\right), \tag{4.21}$$

where $(r,s) \in \{(1,2),(1,3),(2,3)\}$ and $B_k(t)$ are cubic regression splines.

As indicated in (4.21), the hazards are modelled with 7 knots each, hence the total number of parameters is 23. On a computer with 30 cores and using parallel computing (Analytics and Weston, 2014a,b) , model 4.21 takes less than 22 minutes to be estimated.

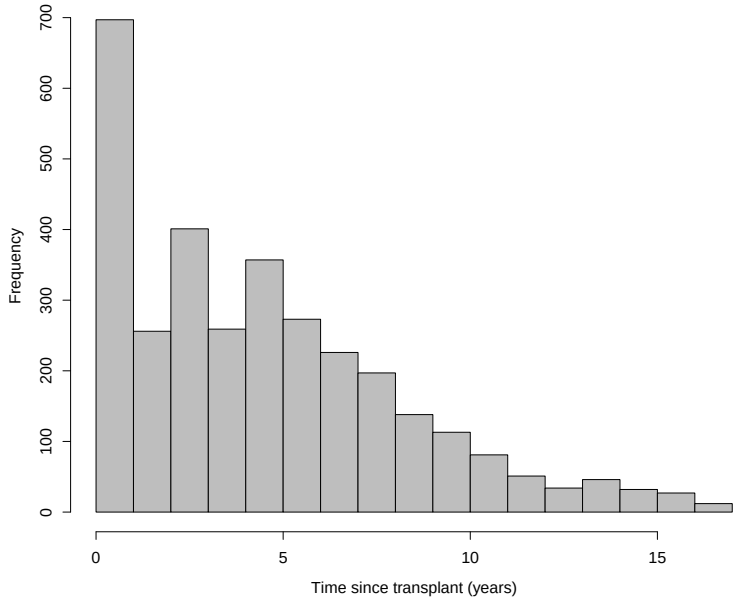

Figure 4.8: Histogram of time since transplant in the CAV data

The vector of smoothing parameters is $\boldsymbol{\lambda} = (\lambda_{12}, \lambda_{13}, \lambda_{23})^\top$. This model has $AIC = 2931.715$, $-2\ell(\boldsymbol{\theta}) = 2903.483$, and $df = 13.846$. Compared to model (2.28), which has $AIC = 2932.953$, $-2\ell(\boldsymbol{\theta}) = 2924.953$, and 8 parameters, model (4.21) is more complex, considerably improves the likelihood, and slightly improves the AIC.

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 4.9. The vector of smoothing parameters is estimated at $\widehat{\boldsymbol{\lambda}} = (14.292, 55.036, 743629)^\top$. 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. That is because data become scarce after that time.

For the parametric part of the model, $\widehat{\beta}_1 = 0.014$ (0.004) and $\widehat{\beta}_2 = 0.2$ (0.096) indicating that donor age and IHD increase the risks of disease progression and

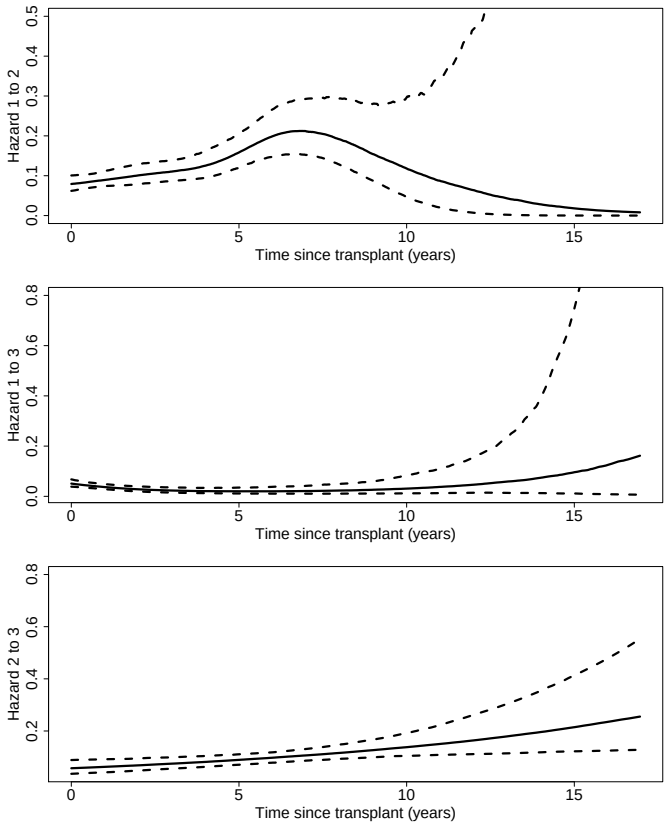

Figure 4.9: Estimated smooth hazards for subjects with IHD and with donor age of 26 (solid lines), with 95% confidence intervals (dashed lines)

death. The model (2.28) has very similar estimates for covariates *dage* and *IHD*, 0.018 and 0.277, respectively. Hence, as in the application to the CAV data, the baseline hazard specification is robust with relation to specification of time-dependency.

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

$$\widehat{\mathbf{P}}(0,5) = \begin{pmatrix} 0.502\ (0.446, 0.556) & 0.318\ (0.274, 0.365) & 0.180\ (0.149, 0.212) \\ 0 & 0.699\ (0.616, 0.773) & 0.301\ (0.227, 0.384) \\ 0 & 0 & 1 \end{pmatrix},$$

with 95% confidence interval (in brackets) obtained using $b = 1000$ simulations as

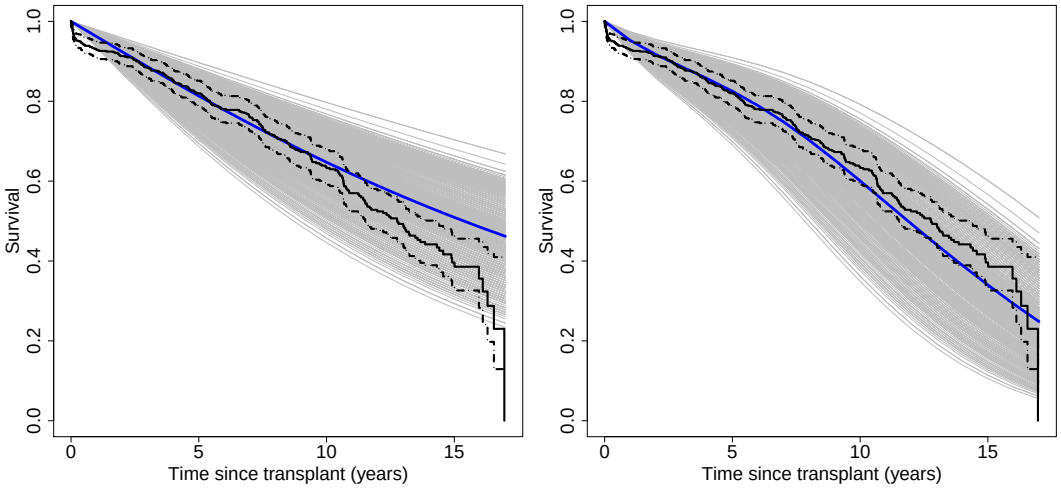

Figure 4.10: Comparison of model-based survival with Kaplan-Meier curves for the Gompertz model (left-hand side) and spline model (right-hand side). Model-based survival: grey lines for individuals, blue lines for the mean of the individual curves. Kaplan-Meier in black lines with 95% confidence intervals

in Section 2.5. 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.

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 4.10 depicts baseline-specific survival as estimated by the models (2.28) and (4.21) (on the left and right hand side, respectively) and as described by the Kaplan-Meier curves. For the Gompertz model in (2.28), 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.

4.5 Discussion

This chapter presents a practical and unifying framework for estimating multi-state models with splines for interval-censored data. This novel methodology improves both accuracy of the parameter estimates and computational speed, if compared to grid search methods such as the one presented in Chapter 3. The new estimation procedure is made possible by rewriting the optimisation problem using a penalised general likelihood estimation (Marra et al., 2017).

The method is illustrated with two applications. We aim to illustrate the feasibility of the method and its usage for flexible time-dependent modelling. In particular, the application to the OCTO data shows that the method is able to efficiently estimate the hazard functions for a complex multi-state processes with a backward transition. The application to the CAV data illustrates the flexibility of the method to model nonlinear time dependency.

The small simulation study and application show the importance of this method for flexible modelling of time-dependent processes. Even though an assessment of bias is not possible due to the small number of replications, the simulation study shows that the method can recover nonlinear, log-liner and linear hazards. Furthermore, we illustrate with an application that the method can give insightful information on the functional form underlying the hazards.

The automatic smoothing parameters estimation as described in Marra et al. (2017) requires the Hessian or the Fisher information for estimation. We show through a simulation and applications 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.

The msm package (Jackson, 2011) is 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 also approached by using a piecewise-constant approximation to the hazards. Therefore, this research can also be seen as a generalisation of the msm package, which allows for flexible modelling of the time-dependency.

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 4.10, the multi-state model with splines can improve considerably model fit by allowing for flexible hazards specification.

Chapter 5

Conclusions and future work

Multi-state models are commonly used in medical research where patients' health status over time is of interest. In the presence of interval censoring, models can be formulated in a Markov processes framework. Models are specified through proportional hazards functions. For time-dependent processes, deciding on a reasonable shape underlying the hazards is not straightforward. Semi-parametric hazards specifications, which allow for flexible modelling of time dependency, are alternatives to parametric models. However, estimating multi-state models with splines is challenging as the problem involves multiple smoothing parameters selection.

This thesis presented novel penalised maximum likelihood methods to estimate multi-state models with splines for interval-censored data. In particular, we have focused on estimating semi-parametric hazards functions with splines. Special attention has been given to cubic regression splines and P-splines, but in principle any other splines basis could have been employed. The principal contribution of this thesis has been to develop an unifying and efficient framework to estimate semi-parametric multi-state models. This new method could be used to model nonlinear time-dependency in multi-state processes or to check parametric model assumptions.

In Chapter 1, we have discussed various multi-state processes and some of their special features, such as model structure and censoring. We also provided a literature review of the existing methods to analyse these processes. We concluded that current methods cannot fully address the problem of estimating multi-state models

with splines.

Chapter 2 then explored a standard method for fitting parametric multi-state models. This method uses a scoring algorithm to maximise the log-likelihood function, for a given piecewise-constant approximation to the hazards. To assess model fit, we used the comparison of the overall model survival function with the Kaplan-Meier empirical estimates. The application of these methods to the OCTO data showed that Gompertz hazards specification seems to lead to a reasonably good model fit. However, such restrictive model assumptions for analysing the CAV data resulted in lack of fit. We then concluded that a method to estimate more flexible multi-state models should be developed to deal with time-dependency.

For this purpose, in Chapter 3, we developed a penalised maximum likelihood method to estimate flexible multi-state models with splines. Specifically, the method allows for parametric and semi-parametric hazards specifications. We showed that a piecewise-constant approximation can be used to calculate the transition probabilities for the likelihood function, and then a scoring algorithm can be used to optimise the penalised likelihood function. The main contribution of this method has been to provide a general framework to specify and estimate multi-state models with splines. We then applied the method to the ELSA data to investigate the decline of cognitive function in older population. Even though time of death is is rounded to the nearest integer, we assumed that time of transition into the dead state is known exactly. This application showed that the risks of moving forward across cognitive states are increasing. The method, however, is based on a grid search for estimating the optimal amount of smoothing. Such approach can become impractical for applications where several hazards are modelled with splines.

In Chapter 4, we extended the method presented in Chapter 3 by developing an efficient algorithm based on penalised maximum likelihood and automatic search of the smoothing parameters to estimate multi-state models for interval-censored data. This method improves both estimation time and, by providing more precise smoothing parameter estimates, accuracy of the model parameter estimates. The method is a general tool that, in principle, allows for flexible modelling of multi-

state processes with any structure, and it is the main contribution of this thesis.

The method has been illustrated with analysis of two data sets and a simulation study. The application to the OCTO data showed that the method can elegantly estimate the smooth hazards of a considerably complicated process with a backward transition. This allowed us to verify that a Gompertz hazard specification can be a reasonable model assumption. In relation to the application to the CAV data, the estimated smooth hazards showed that the risks of moving between states follow a nonlinear trend over time. We verified through a model diagnostic that such flexible hazards specification with splines improves the model fitting for the CAV data. Therefore, the method enabled us to improve the knowledge of the process of interest. Finally, the simulation study allowed us to verify that the method can recover non-linear and linear hazards functions.

The piecewise-constant approximation provides a general framework to calculate transition probabilities of any multi-state process. However, the method requires a much greater number of numerical evaluations as computation of eigenvalues decomposition is required for all pair of observation times. Alternatively, the method presented in Titman (2011) could also be used to calculate transition probabilities. Theoretical properties such as conditions for consistency and identifiability should be studied.

The estimation time is improved by using parallel computing to calculate the score vector and Fisher information matrix. The time to estimate model (4.21) with a computer with 30 cores is just below 22 minutes. Computational time could be further improved by using the package in Rcpp (Eddelbuettel et al., 2011).

To conclude, the methods presented in this thesis can be further extended to estimate multi-state models for different data structures. One possible extension is to allow for misclassification of states (Jackson et al., 2003). This poses extra difficulty for estimation as computation of the gradient or information is more challenging (Lystig and Hughes, 2002). For processes with time-varying covariates, the hazards are specified in terms of functions of the covariates. Similarly to the specification of the baseline hazards, the functional forms underlying time-varying covariates are

often unknown. Therefore, these functions could be specified by smooth functions. Semiparametric hazards specification with splines of the covariate effects can be useful to provide insightful information about the process of interest. The work by Joly and Commenges (1999) on penalised semi-Markov models could be extended in two ways. First, an automatic smoothing parameter selection method can be used to improve precision of parameter estimation. Second, the method could be extended for penalized semi-Markov models with backward transitions.

Appendix A

Code for the R software

The method developed in this thesis are implemented in the software R (R Core Team, 2016). In this appendix we provide the code for fitting the three-state model for the CAV data, as defined in Section 4.4.2.

```
#It is necessary to load the following libraries

library(msm)
library(splines)
library(mgcv)
library(flexsurv)


#The CAV data is assigned to the variable ``Data''.
#The following is the CAV data for an individual.

   PTNUM    years state dage pdiag
1 100002 0.000000     1   21     1
2 100002 1.002740     1   21     1
3 100002 2.002740     2   21     1
4 100002 3.093151     2   21     1
5 100002 4.000000     2   21     1
6 100002 4.997260     2   21     1
7 100002 5.854795     3   21     1

# The following construct the smoothing basis.
# The argument ``bs'' represent spline basis.
#In this case, ``cr'' stands for cubic regression splines.

smoother =  smooth.construct(s(years, bs = "cr"), Data, NULL)

#Get the design matrix

X = smoother$X
```

```
#Get the dimension of the smoother

n.dim = smoother$bs.dim

#Get the penalty matrix
pen = output <- matrix(unlist(smoother$S), ncol = n.dim, byrow = TRUE)

#Define the number of parameter for each transition and
#the number of covariates.

npar12 = n.dim
npar13 = n.dim
npar23 = n.dim
ncov = 2

#This is useful for coding
n1 = npar12
n2 = n1 + npar13
n3 = n2 + npar23
n4 = n3 + ncov
npar = n4

#Vector of initial values
par.0 = c(rep(-3, n3), rep(0,2))

#number of states and dead state
s.number = 3
dead.state = 3

#Split the data into individuals
n = split(Data, Data$PTNUM)
N = length(n)

#Define elements for the scoring
iter=1
max.iter=100
diff= rep(Inf,length(par))

digits = 3
count = 0
Max = 1000
abstol = 1e-6

gamma = c(2, 2, 2)
sp = exp(gamma)


#The next function is the generator matrix and its eigen decomposition
```

```
#Data.Ind.i is the data for ith individual
#X.i smoother for ith individual
#j is the jth observation

Q.matrix = function(par, Data.Ind.i, j, X.i){
        dage = Data.Ind.i[j-1,4]
        pdiag = Data.Ind.i[j-1,5]
        Q = matrix(0,s.number,s.number)
        Q[1,2] = exp(X.i[j-1,]%*%par[1:n1] + par[n3+1]*dage
        + par[n3+2]*pdiag)
        Q[1,3] = exp(X.i[j-1,]%*%par[(n1+1):n2] + par[n3+1]*dage
        + par[n3+2]*pdiag)
        Q[1,1] = -(Q[1,2] + Q[1,3])
        Q[2,3] = exp(X.i[j-1,]%*%par[(n2+1):n3] + par[n3+1]*dage
        + par[n3+2]*pdiag)
        Q[2,2] = -Q[2,3]
        r = eigen(Q, symmetric = FALSE)
        A = r$vectors
        d = r$values
        inv.A = solve(A)
        Q.result = list("A" = A, "d" = d, "inv" = inv.A,
        "Q" = Q)
        return(Q.result)
}

#Function for calculating the probability transition for a follow-up
P = function(j, Data.Ind.i, A, d, inv.A){
        t = exp(d*(Data.Ind.i[j,2] - Data.Ind.i[j-1,2]))
        P = A%*%diag(t)%*%inv.A
        return(P)
}

#Function that returns  a likelihood contribution for a
#interval-censored observation time
Pt = function(j, Data.Ind.i, Q){
        r = eigen(Q, symmetric = FALSE)
        A = r$vectors
        x = r$values
        t = exp(x*(Data.Ind.i[j,2] - Data.Ind.i[j-1,2]))
        pt.aux = A%*%diag(t)%*%solve(A)
        pt = pt.aux[Data.Ind.i[j-1,3], Data.Ind.i[j,3]]
        return(pt)
}

#Function that returns  a likelihood contribution for
#death times

Ps = function(s, Data.Ind.i, Q){
        m = nrow(Data.Ind.i)
        r = eigen(Q, symmetric = FALSE)
```

```
        A = r$vectors
        x = r$values
        t = exp(x*(Data.Ind.i[m,2] - Data.Ind.i[m-1,2]))
        ps.aux = A%*%diag(t)%*%solve(A)
        ps = ps.aux[Data.Ind.i[m-1,3], s]*Q[s,s.number]
        return(ps)
}

#Function to calculate the trace of a function
trace = function(A){
        s = 0
        for (i in 1:(nrow(A))){
                s = s + A[i,i]
        }
        return(s)
}

#Matrix square root
square.root = function(S){
        d = eigen(S, symmetric = TRUE)
        rS = d$vectors%*%diag(d$values^0.5)%*%t(d$vectors)
}

#Derivative of the Q matrix
deriv.Q = function(par,Data.Ind.i, j, u, Q, X.i){
        dage = Data.Ind.i[j-1,4]
        pdiag = Data.Ind.i[j-1,5]
        dQ = matrix(0, s.number, s.number)
        if(u<=n1){
                dQ[1,2] = X.i[j-1,u]*Q[1,2]
                dQ[1,1] = -dQ[1,2]
        }
        if(u>n1 & u<=n2){
                dQ[1,3] = X.i[j-1,u-n1]*Q[1,3]
                dQ[1,1] = -dQ[1,3]
        }
        if(u>n2 & u<=n3){
                dQ[2,3] = X.i[j-1,u-n2]*Q[2,3]
                dQ[2,2] = -dQ[2,3]
        }
        if(u==n3+1){
                dQ[1,2] = dage*Q[1,2]
                dQ[1,3] = dage*Q[1,3]
                dQ[1,1] = -(dQ[1,2]+dQ[1,3])
                dQ[2,3] = dage*Q[2,3]
                dQ[2,2] = -dQ[2,3]

        }
        if(u==n3+2){
                dQ[1,2] = pdiag*Q[1,2]
```

```
                dQ[1,3] = pdiag*Q[1,3]
                dQ[1,1] = -(dQ[1,2]+dQ[1,3])
                dQ[2,3] = pdiag*Q[2,3]
                dQ[2,2] = -dQ[2,3]

        }
        return(dQ)
}

#Derivative of the probability transition matrix
deriv.P = function(j, Data.Ind.i, dQ.u, s.number, A, d, inv.A){
        t = (Data.Ind.i[j,2] - Data.Ind.i[j-1,2])
        G.u = inv.A%*%dQ.u%*%A
        V.u = matrix(0,s.number, s.number)
        for(l in 1:s.number){
                for(m in 1:s.number){
                        if(l==m){
                                V.u[l,l] = G.u[l,l]*t*exp(d[l]*t)
                        }else{
                                V.u[l,m] = G.u[l,m]*(exp(d[l]*t)
                                - exp(d[m]*t))/(d[l] - d[m])
                        }
                }
        }
        dP.u = A%*%V.u%*%inv.A
        return(dP.u)
}

#Function to calculate the gradient and the M matrix
scoring.msm = function(N, par){
        M = matrix(0,length(par), length(par))
        S = matrix(0,length(par),1)
        for(u in 1:length(par)){
                for(v in 1:length(par)){
                        if(u <= v){
                                aux = scoring.uv(u,v,N,par)
                                M[v,u] = aux[[2]]
                                M[u,v] = aux[[2]]
                                if(u==v){
                                        S[u] = aux[[1]]
                                }
                        }

                }
        }
        result = list("S" = S, "M" = M)
        return(result)
}
```

```
#Calculates the u entry of the gradient
#and the uv entry of the M matrix
scoring.uv = function(u, v, N, par){
        M.uv = 0
        score.u.aux = 0
        for(i in 1:N){
                Data.Ind.i = n[[i]]
                X.i = X[(1:nrow(Data.Ind.i)),] #get data associated to subject i
                X = X[-(1:nrow(Data.Ind.i)),] # Delete data associated to subject i
                for(j in 2:nrow(Data.Ind.i)){
                        decomp.Q = Q.matrix(par, Data.Ind.i, j, X.i)
                        A = decomp.Q$A
                        d = decomp.Q$d
                        inv.A = decomp.Q$inv
                        Q = decomp.Q$Q
                        P.aux = P(j, Data.Ind.i, A, d, inv.A)
                        dQ.u = deriv.Q(par, Data.Ind.i, j, u, Q, X.i)
                        dQ.v = deriv.Q(par, Data.Ind.i, j, v, Q, X.i)
                        dP.u = deriv.P(j, Data.Ind.i, dQ.u, s.number, A, d, inv.A)
                        dP.v = deriv.P(j, Data.Ind.i, dQ.v, s.number, A, d, inv.A)
                        if(Data.Ind.i[j,3]!=s.number){
                                M.uv = M.uv + (1/(P.aux[Data.Ind.i[j-1,3], Data.Ind.i[j,3]])^2)
                                *dP.u[Data.Ind.i[j-1,3], Data.Ind.i[j,3]]
                                *dP.v[Data.Ind.i[j-1,3], Data.Ind.i[j,3]]
                                score.u.aux = score.u.aux
                                + (1/P.aux[Data.Ind.i[j-1,3], Data.Ind.i[j,3]])
                                *dP.u[Data.Ind.i[j-1,3], Data.Ind.i[j,3]]
                        }else{
                                denom = 0
```

```
                                num.u = 0
                                num.v = 0
                                for(s in 1:(s.number-1)){
                                        denom = denom + P.aux[Data.Ind.i[j-1,3], s]*Q[s, s.number]
                                        num.u = num.u + dP.u[Data.Ind.i[j-1,3], s]*Q[s, s.number]
                                        + P.aux[Data.Ind.i[j-1,3], s]*dQ.u[s, s.number]
                                        num.v = num.v + dP.v[Data.Ind.i[j-1,3], s]*Q[s, s.number]
                                        + P.aux[Data.Ind.i[j-1,3], s]*dQ.v[s, s.number]
                                }
                                M.uv = M.uv + (1/denom^2)*num.u*num.v
                                score.u.aux = score.u.aux + num.u/denom
                        }
                }
        }
        result = list("score.u" = score.u.aux, "M.uv" = M.uv)
        return(result)
}
```

```
#Set up the penalty matrix
spl.S = function(n1, n2, n3, n4, npar,pen){
        Pen = list()
        Pen[[1]] = Pen[[2]] = Pen[[3]] = Pen[[4]]
        = matrix(0, npar, npar)
        Pen[[1]][1:n1, 1:n1] = pen
        Pen[[2]][(n1+1):n2,(n1+1):n2] = pen
        Pen[[3]][(n2+1):n3,(n2+1):n3] = pen
        return(Pen)
}

#Estimate the model parameter for given
#smoothing parameters
estimate.parameters = function(par.0, sp){
        count=0
        Pen = spl.S(n1, n2, n3, n4, npar,pen)
        Pen = sp[1]*Pen[[1]] + sp[2]*Pen[[2]] + sp[3]*Pen[[3]]
        Max.inner = 50
        while(count < Max.inner){
                aux = scoring.msm(N,par.0)
                S = aux[[1]]
                Sp = S - Pen%*%par.0
                I = aux[[2]]
                Ip = I + Pen
                sqrt.I = square.root(I)
                inv.sqrt.I = solve(sqrt.I)
                e = inv.sqrt.I%*%S
                z = sqrt.I%*%par.0 + e
                inv.pen.I = solve(I + Pen)
                par.h = inv.pen.I%*%sqrt.I%*%z
                count = count + 1
                if(max(abs(par.0 - par.h)) < abstol) break
                par.0 = par.h
        }
        est.parameters.result = list("par" = par.h, "I" = I,
        "Ip"=Ip, "inv.pen.I" = inv.pen.I, "sqrt.I" = sqrt.I,
        "z" = z, "count"=count)
        return(est.parameters.result)
}
```

```
#Set up the UBRE
#gamma represents log(sp)
UBRE = function(gamma){
        Pen = spl.S(n1, n2, n3, n4, npar, pen)
        sp = exp(gamma)
        Pen = sp[1]*Pen[[1]] + sp[2]*Pen[[2]] + sp[3]*Pen[[3]]
        A = sqrt.I%*%solve(I + Pen)%*%sqrt.I
        ubre = t(z - A%*%z)%*%(z - A%*%z) + 2*trace(A) - npar
        return(ubre)
}

#Minimise the UBRE
estimate.lambda = function(gamma){
        lambda.update = optim(gamma, UBRE, method = "BFGS",
        control = list(maxit = 100000), hessian = TRUE)
        lambda = lambda.update$par
        return(lambda)
}

#The following while runs until the convergence criteria is achieved
# It is the summary of the scoring algorithm
while(count < Max){
        find.est = estimate.parameters(par.0, sp)
        par = find.est[[1]]
        I = find.est[[2]]
        inv.pen.I = find.est[[4]]
        sqrt.I = find.est[[5]]
        z = find.est[[6]]
        count = count + 1
        if(max(abs(par - par.0)) < abstol) break
        est.gamma = estimate.lambda(gamma)
        gamma = est.gamma
        sp = exp(gamma)
        par.0 = par
}
```

Bibliography

E. L. Abner, R. J. Kryscio, G. E. Cooper, D. W. Fardo, G. A. Jicha, M. S. Mendiondo, P. T. Nelson, C. D. Smith, L. J. Van Eldik, L. Wan, et al. Mild cognitive impairment: Statistical models of transition using longitudinal clinical data. *International Journal of Alzheimer's Disease*, 2012:291920, 2012.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

R. Analytics and S. Weston. *doParallel: Foreach parallel adaptor for the parallel package*, 2014a. URL `http://CRAN.R-project.org/package=doParallel`. R package version 1.0.8.

R. Analytics and S. Weston. *foreach: Foreach looping construct for R*, 2014b. URL `http://CRAN.R-project.org/package=foreach`. R package version 1.4.2.

P. K. Andersen and N. Keiding. Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115, 2002.

E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' guide. Third Edition.* Philadelphia: SIAM, 1999.

R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006.

D. Collett. *Modelling survival data in medical research*. CRC press, 2015.

D. Commenges. Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research*, 11(2):167–182, 2002.

D. Commenges, P. Joly, A. Gégout-Petit, and B. Liquet. Choice between semi-parametric estimators of markov and non-markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics*, 34(1):33–52, 2007.

R. J. Cook and J. F. Lawless. Analysis of repeated events. *Statistical Methods in Medical Research*, 11(2):141–166, 2002.

D. R. Cox. *The theory of stochastic processes*. Routledge, 2017.

P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

C. De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

D. Eddelbuettel, R. François, J. Allaire, K. Ushey, Q. Kou, N. Russel, J. Chambers, and D. Bates. Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

P. H. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–102, 1996.

L. Fahrmeir and A. Klinger. A nonparametric multiplicative hazard model for event history analysis. *Biometrika*, 85(3):581–592, 1998.

M. F. Folstein, S. E. Folstein, and P. R. McHugh. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975.

R. Gentleman, J. Lawless, J. Lindsey, and P. Yan. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13(8):805–821, 1994.

R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951, 1992.

C. Gu and G. Wahba. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computing*, 12(2):383–398, 1991.

R. A. Hubbard, L. Inoue, and J. Fann. Modeling nonhomogeneous markov processes via time transformation. *Biometrics*, 64(3):843–850, 2008.

C. H. Jackson. Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, 38(8):1–29, 2011.

C. H. Jackson. Flexsurv: A platform for parametric survival modelling in R. *Journal of Statistical Software*, 70(8):1–33, 2016.

C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, 2003.

P. Joly and D. Commenges. A penalized likelihood approach for a progressive three-state model with censored and truncated Data: Application to AIDS. *Biometrics*, 55(3):887–890, 1999.

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

P. Joly, C. Durand, C. Helmer, and D. Commenges. Estimating life expectancy of demented and institutionalized subjects from interval-censored observations of a multi-state model. *Statistical Modelling*, 9(4):345–360, 2009.

J. Kalbfleisch and J. F. Lawless. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392):863–871, 1985.

R. Kay. A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 80:855–865, 1986.

T. Kneib and A. Hennerfeind. Bayesian semi-parametric multi-state models. *Statistical Modelling*, 8(2):169–198, 2008.

T. C. Lystig and J. P. Hughes. Exact computation of the observed information matrix for hidden markov models. *Journal of Computational and Graphical Statistics*, 11(3):678–689, 2002.

R. J. Machado and A. van den Hout. Flexible multistate models for interval-censored data: Specification, estimation, and an application to ageing research. *Statistics in Medicine*, 37(10):1636–1649, 2018.

R. J. M. Machado, A. Van den Hout, and G. Marra. Penalised maximum likelihood estimation in multistate models for interval-censored data. *arXiv preprint arXiv:1801.06375*, 2018.

R. E. Marioni, M. J. Valenzuela, A. Van den Hout, C. Brayne, F. E. Matthews, et al. Active cognitive lifestyle is associated with positive cognitive health transitions and compression of morbidity from age sixty-five. *PLoS One*, 7(12):e50940, 2012.

G. Marra, R. Radice, T. Bärnighausen, S. N. Wood, and M. E. McGovern. A Simultaneous Equation Approach to Estimating HIV Prevalence With Nonignorable Missing Responses. *Journal of the American Statistical Association*, 112(518): 484–496, 2017.

C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1):3–49, 2003.

R. Z. Omar, N. Stallard, and J. Whitehead. A parametric multistate model for the analysis of carcinogenicity experiments. *Lifetime Data Analysis*, 1(4):327–346, 1995.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

R. Radice, G. Marra, and M. Wojtyś. Copula regression spline models for binary outcomes. *Statistics and Computing*, 26(5):981–995, 2016.

A. Robitaille, A. van den Hout, R. J. Machado, D. A. Bennett, I. Čukić, I. J. Deary, S. M. Hofer, E. O. Hoogendijk, M. Huisman, B. Johansson, et al. 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*, 2018.

P. Royston and M. K. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21 (15):2175–2197, 2002.

D. Ruppert, M. P. Wand, and R. J. Carroll. Semiparametric regression. cambridge series in statistical and probabilistic mathematics 12. *Cambridge: Cambridge Univ. Press. Mathematical Reviews (MathSciNet): MR1998720*, 2003.

G. A. Satten and I. M. Longini. Markov chains with measurement error: estimating thetrue'course of a marker of the progression of human immunodeficiency virus disease. *Applied Statistics*, pages 275–309, 1996.

H. Sennhenn-Reulen and T. Kneib. Structured fusion lasso penalized multi-state models. *Statistics in Medicine*, 35(25):4637–4659, 2016.

L. D. Sharples, C. H. Jackson, J. Parameshwar, J. Wallwork, and S. R. Large. Diagnostic accuracy of coronary angiography and risk factors for post–heart-transplant cardiac allograft vasculopathy. *Transplantation*, 76(4):679–682, 2003.

A. C. Titman. *Model diagnostics in multi-state models of biological systems*. PhD thesis, University of Cambridge, 2008.

A. C. Titman. Flexible nonhomogeneous Markov models for panel observed data. *Biometrics*, 67(3):780–787, 2011.

A. C. Titman and L. D. Sharples. Model diagnostics for multi-state models. *Statistical Methods in Medical Research*, 19(6):621–651, 2010.

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

A. Van den Hout and F. E. Matthews. Multi-state analysis of cognitive ability data: A piecewise-constant model and a Weibull model. *Statistics in Medicine*, 27(26): 5440–5455, 2008a.

A. Van den Hout and F. E. Matthews. A piecewise-constant markov model and the effects of study design on the estimation of life expectancies in health and ill health. *Statistical Methods in Medical Research*, 2008b.

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

S. N. Wood. Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 413–428, 2000.

S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99 (467), 2004.

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