THE JOINT MODELS FOR NON-LINEAR LONGITUDINAL AND TIME-TO-EVENT DATA USING PENALIZED SPLINES

A dissertation submitted for the degree of Doctor of Philosophy (Statistics)

by

Huong Thi Thu Pham

College of Science and Engineering

Flinders University

6^{th} April 2018

Contents

List of Figures

List of Tables

List of Abbreviations

ACF	Autocorrelation function
AIC	Akaike's information criterion
AIDS	Acquired immunodeficiency syndrome
BIC	Bayesian information criterion
BLUPs	Best linear unbiased predictors
CIs	confidence intervals
CrIs	Credible intervals
DPM	Dirichlet process model
EM	Expectation maximization
ECM	Expectation conditional maximization
GS	Gibbs sampler
LMEs	Linear mixed effects models
MCMC	Markov chain Monte Carlo
MH	Metropolis Hastings
MSE	Mean Square Error
Psrf	Potential scale reduction factors
SD	Standard deviation
SE	Standard Error
$\mathcal{G}$	Gamma distribution
$\mathcal{IG}$	Inverse gamma distribution
$\mathcal{IW}$	Inverse Wishart distribution
$\mathcal{W}$	Wishart distribution
$\mathcal{MVN}$	Multivariate normal distribution
$\mathcal{N}$	Normal distribution
$\mathcal{U}$	Uniform distribution
$\mathcal{UVN}$	Univariate normal distribution

Summary

Joint models for longitudinal and time-to-event data have been applied in many different fields of statistics and clinical studies. My interest is in modelling the relationship between event time outcomes and internal time-dependent covariates. In practice, the longitudinal responses often show non-linear and fluctuated curves. Therefore, the main aim of this thesis is to use penalized splines with a truncated polynomial basis to parameterise the non-linear longitudinal process. Then, the linear mixed effects model is applied to subject-specific curves and to control the smoothing. The association between the dropout process and longitudinal outcomes is modeled through a proportional hazard model. Two types of baseline risk functions are considered, namely a Gompertz distribution and a piecewise constant model. The resulting models are referred to as penalized spline joint models; an extension of the standard joint models. The expectation conditional maximization (ECM) algorithm is applied to estimate the parameters in the proposed models. To validate the proposed algorithm, extensive simulation studies were implemented followed by a case study. Simulation studies show that the penalized spline joint models improve the existing standard joint models.

The main difficulty that the penalized spline joint models have to face with is the computational problem. The requirement for numerical integration becomes severe when the dimension of random effects increases. In this thesis, a modified two-stage approach has been proposed to estimate the parameters in joint models. This approach not only improves a previous two-stage approach but also allows for the application of extended joint models with a high dimension of random effects in the longitudinal submodel. In particular, in the first stage, the linear mixed effects models (LMEs) and best linear unbiased predictors (BLUPs) are applied to estimate parameters in the longitudinal submodel. Then, in the second stage, an approximation of the fully joint log-likelihood is proposed using the estimated values of these parameters from the longitudinal submodel. The survival parameters are estimated by maximizing the approximation of the fully joint

log-likelihood. Simulation studies show that the modified two-stage approach performs well, especially when the dimension of the random effects in the penalized splines joint models increases.

Finally, a Bayesian approach is applied to estimate the parameters in the penalized splines joint models. This approach provides alternative ways to infer the uncertainties of the parameters in the penalized splines joint models. Moreover, this approach can avoid approximations resulting from calculating multiple integrals in the frequentist approach. The Markov chain Monte Carlo (MCMC) algorithm is proposed containing the Gibbs sampler (GS) and Metropolis Hastings (MH) algorithms to sample for the target conditional posterior distributions. Extensive simulation studies were implemented to validate the proposed algorithm. In addition, the prior sensitivity analysis for the baseline hazard rate and association parameters is performed through simulation studies and a case study. The results show that the fully Bayesian approach produces reliable estimates and complete inferences for the parameters in the penalized splines joint models.

Declaration

Declaration

I declare that:

This thesis is my own work and does not incorporate any material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma except where referenced or acknowledged.

To the best of my knowledge, this thesis does not contain, without acknowledgement, any material previously published or written by another person.

Huong Thi Thu Pham

April 2018

Publications

This thesis was completed under the supervison of Dr Darfiana Nur, Associate Professor Alan Branford and Associate Professor Murk Bottema. Shortened version of the three main chapters in this thesis have been submitted to statistical journals. The list is as follows:

Published Journal Articles

P. Huong, D. Nur, and A. Branford. Penalized spline joint models for longitudinal and time-to-event data. Communication in Statistics Theory and Methods, 2016.

[DOI: 10.1080/03610926.2016.1235195]

Submitted Manuscripts

P. Huong, D. Nur, and A. Branford. A modified two-stage approach for joint modelling of longitudinal and time-to-event data, Computational Statistics.

P. Huong, D. Nur, and A. Branford. A prior sensitivity analysis for joint modelling of longitudinal and time-to-event data, Journal of Statistical Computation and Simulation.

Acknowledgements

The chance to come to Australia and study at Flinders University is one of the most prominent events in my life. I have experienced and learned a lot of new things in my international student life at Flinders University. In this journey, I have received tremendous support from my supervisors, friends and family to overcome the hardships in research as well as in daily life.

I would like to express my sincere gratitude to all of my supervisors. Firstly, I would like to thank my main supervisor Doctor Darfiana Nur. Darfiana has given great support to me both academically and psychologically at the time I was confused and not confident in completing this thesis. Thank you very much for your encouragement and help in this long journey. Secondly, I am grateful to Associate Professor Alan J. Branford. Alan introduced me to the subject of survival analysis and the work of Dimitri Rizopoulos. With this initial help, I became interested in the joint modelling framework and came up with the ideas to contribute to this field. In addition, my special thanks is offered to Associate Professor Murk Bottema for his useful suggestions and support. I was very happy to be his student and to sit in front of his office. Finally, to Professor Jerzy Filar and Doctor Ray Booth for sharing their knowledge of doing research and editing my work.

My PhD journey was more pleasant and enjoyable when I received the support from my family and my friends. I also would like to thank the Vietnamese student association at Flinders University for their support and encouragement during all these years.

Last but not least, I wish to thank the Australian Award Scholarship for the financial support and the staff of ISSU for helping international students at Flinders University. The scholarship gave me a great chance to gain good knowledge about statistical analysis and to complete this thesis.

Chapter 1

Introduction

In follow-up type of studies, there are different types of response variables collected for each individual. They are longitudinal outcomes which are measured on each subject repeatedly, and the time when the subject meets an event of particular interest. There are many research questions focusing on the association between longitudinal data and survival time in clinical, epidemiological and educational studies. In many clinical studies, the researchers want to evaluate the impact of biomarkers for their prognostic capabilities on survival time outcomes. Tsiatis et al. (1995) investigated the association between the number of CD4-lymphocyte and the time to death in an acquired immune deficiency syndrome (AIDS) study. The link between serum bilirubin level and survival time was investigated in liver cirrhosis studies (Rizopoulos, 2011; Ding and Wang, 2008). In addition, there has been interest in the interrelation between these two types of data in other fields. For instance, the environmental factors or seasonal patterns may be associated with the occurrence of some types of diseases such as asthma or depression (Rizopoulos, 2012; Kalbfleisch and Prentice, 2002).

Joint models aim to measure the association between survival time and longitudinal responses. These models can be used to better estimate the survival and longitudinal processes as well as evaluating their association. There are different types of longitudinal covariates and there is a demand on modelling survival time and trajectory for each individual. Therefore, flexible joint models are introduced to suit each type of longitudinal covariate and parameterize individual curves (Cox, 1972, 1975; Andersen et al., 1993; Rizopoulos, 2012; Tsiatis and Davidian, 2004). In addition, different approaches and techniques need to be considered to estimate parameters for joint models (Cox and

Hinkley, 1979; Tsiatis and Davidian, 2001; Rizopoulos, 2011; Ibrahim et al., 2005; Gould et al., 2014).

Cox (1972, 1975) introduced joint models using proportional hazard models. The Cox model has been, and remains, a very popular joint model to deal with time-independent covariates using a partial likelihood approach. However, the Cox model contains many disadvantages for handling time-dependent covariates (Cox, 1972). Time-dependent covariates are also divided into two types which are external and internal covariates. Cox (1975) extended his method to handle the external longitudinal covariates. These models are known as the extended Cox models, which also use the partial likelihood approach for estimation (Cox, 1975; Cox and Hinkley, 1979; Cox and Oakes, 1984; Andersen et al., 1993).

Another category of time-dependent covariates is internal longitudinal outcomes, which can be found in many clinical studies. The extended Cox model using a partial likelihood approach can cause large biases and poor coverage properties for handling internal covariates Sweeting and Thompson (2011); Tsiatis and Davidian (2004). Rizopoulos (2012) proposed standard joint models postulating from the proportional hazard model. He used the full likelihood approach to estimate the parameters in the joint models. This approach performs acceptably better for handling internal covariates compared to the Cox model and the extended Cox model (Rizopoulos, 2012; Gould et al., 2014).

In the full likelihood approach, the whole history of biomarkers influences the survival function. Thus, it is important to obtain good models for longitudinal data in order to estimate the survival time accurately. Moreover in practice, subject-specific trajectories may show non-linear curves for a long period of measurement. Estimating parameters for standard joint models is often quick and easy. However, they may not fit non-linear longitudinal data and especially cannot handle smoothing. This potential problem can be addressed by proposing an appropriate longitudinal submodel to handle non-linear longitudinal data Gould et al. (2014); Tsiatis and Davidian (2004). In this thesis, we mainly focus on modelling the association between the internal non-linear longitudinal outcomes and event-time outcomes as well as parameter estimation using different approaches.

This thesis introduces penalized spline joint models to handle non-linear longitudinal outcomes in Chapter 3. These models are not only a good fit for non-linear longitudinal data, but can also control the roughness of fit for the individual curves. To estimate the

parameters in these models, the full likelihood approach is applied. Particularly, parameter estimation is obtained by using the expectation conditional maximization (ECM) algorithm. These models can improve the biases and the goodness of fit compared to the standard linear joint models. However, the penalized spline joint models can become complicated quickly when the number of knots in the longitudinal submodel increases. The full likelihood approach can lead to a computational problem for which the algorithm takes a long time to converge.

To deal with this computational problem, in this thesis, a modified two-stage approach is proposed in Chapter 4. We introduce an algorithm to estimate the parameters for the penalized spline joint models. This approach allows the allocation of as many knots as possible to the penalized spline joint models. In addition, this approach not only reduces the time for convergence but also has biases comparable to the full likelihood approach.

Finally, to avoid the approximation from calculating multiple integrals in the frequentist approach, and to quantify uncertainty using a probability density function for the penalized spline joint models, a fully Bayesian approach is applied to the penalized spline joint models in Chapter 5. In this approach, based on the likelihood function, we formulate the joint posterior distribution. The main algorithm using the Metropolis Hastings (MH) and Gibbs sampler (GS) algorithms is proposed to sample the parameters for the penalized spline joint models. In addition, prior sensitivity analysis is performed to confirm the results of the inferences based on different prior distributions of some important parameters in joint models.

In summary, the original contributions of this thesis include:

(i) The introduction of penalized spline joint models for non-linear longitudinal data and time-to-event data; In particular, we implement penalized splines using a truncated polynomial basis for the longitudinal submodel (Section 3.2);

(ii) The three approaches being proposed for estimating parameters for penalized spline joint models namely the ECM full likelihood approach (Section 3.3), the modified two-stage approach (Sections 4.2 and 4.3) and the fully Bayesian approach (Sections 5.3 and 5.4);

(iii) Extensive simulation studies in Sections 3.4, 4.4 and 5.5 to validate the three approaches in (ii);

(iv) Random effects misspecification analysis for the modified two-stage approach (Section 4.5);

(v) A prior sensitivity analysis for the Bayesian approach (Section 5.6);

(vi) The R codes written for the three approaches.

To achieve these aims, this thesis is organized into six chapters as follows: Chapter 1 is this introductory chapter. The background for longitudinal analysis, survival analysis and joint modelling are introduced in Chapter 2. The frequentist and Bayesian approaches for joint models are also reviewed in this chapter. Penalized splines models are proposed in Chapter 3. In this chapter, we also introduce the ECM algorithm and a set of R code written to estimate the parameters in the proposed joint models. The modified two-stage approach is introduced in Chapter 4. In this chapter, a proposed two-stage algorithm is also presented and a set of R code is provided. Intensive simulation studies are conducted to compare with the full likelihood approach. Chapter 5 uses a fully Bayesian approach to estimate parameters in the penalized joint models. The Markov chain Monte Carlo (MCMC) method is applied to sample parameters. Finally, conclusions about the main results obtained in this thesis, remaining problems and future research for joint models are discussed in Chapter 6.

Chapter 2

Literature Review

Longitudinal data and survival data frequently occur together in practice. As an example, in many medical studies, patients' information such as CD4 cell counts, serum bilirubin level, etc, are collected repeatedly to be associated with survival time. Recently, a large number of studies investigate the link between a true potential biomarker and survival time Cox (1972); Tsiatis and Davidian (2001); Rizopoulos (2012); Ding and Wang (2008); Ibrahim et al. (2005). Joint models for longitudinal data and time-to-event data aim to measure the association between the longitudinal marker level and event times. These models can be used to obtain a good fit for the longitudinal process and better prediction for the survival process.

There are two important submodels used to build the joint models. These are the linear mixed-effects model and the relative risk model. In this chapter, the background for longitudinal data analysis is first presented in Section 2.1 followed by survival data analysis in Section 2.2. In particular, linear mixed effects models and penalized spline longitudinal models are reviewed for longitudinal data. Cox and extended Cox models are presented for survival analysis. Furthermore, we review the standard joint models for longitudinal and survival data in the literature that have used a frequentist approach to estimate the parameters in the joint models in Section 2.3. At the end, a Bayesian approach, which can be considered to be an alternative method to estimate the parameters in the joint models, is presented in Section 2.4.

2.1 Longitudinal data analysis

Longitudinal data is correlated data measured repeatedly at different time points. This type of data is commonly found in many different fields of quantitative research, especially

in health sciences. To analyse this type of data, well-fitting models and methods are proposed to be able to make inferences for population means and individual means at specific time points. The analysis also investigates the change of these means over time (Cox and Hinkley, 1979; Singer and Willett, 2003).

Longitudinal data analysis has been long developed in the literature. Hand and Crowder (1996), Verbeke and Molenberghs (2000), Diggle et al. (2002) and Molenberghs and Verbeke (2005) provided overviews of the theory for longitudinal data that focus on multivariate regression models and multivariate analysis of variance. Rao (1997), Fitzmaurice et al. (2004), Gelman and Hill (2007) and McCulloch et al. (2008) showed differences between longitudinal data analysis assuming correlated observations and cross sectional data analysis assuming independent observations. They also presented methods for estimating parameters in different longitudinal regression models. Many modern methods have been developed for analysing data from longitudinal studies and many packages for implementing these methods are available for various software environments(Pinheiro et al., 2014; Bates et al., 2011; Venables and Ripley, 2013; Rice and Wu, 2001).

In longitudinal data regression, subject-specific trajectories can either be linear or non-linear curves. There have been numerous studies that have analysed non-linear longitudinal datasets. The relationship between CD4 cell counts and time in the AIDS dataset (Abrams et al., 1994) showed lightly non-linear curves for five repeated measurements. Many profiles in primary biliary cirrhosis data and liver cirrhosis data showed obviously non-linear serum bilirubin levels and prothrombin indexes in time (Andersen et al., 1993; Murtaugh et al., 1994).

To model subject-specific curves having a non-linear response profile over time, the linear mixed effects models and penalized spline regression models for longitudinal data can be used. Linear mixed effects models are effective in estimating not only the population mean but also the individual trajectories as they change over time. These models were investigated by Hand and Crowder (1996), Verbeke and Molenberghs (2000), Fitzmaurice et al. (2004), Ruppert et al. (2009), Jiang (2010), McCulloch and Neuhaus (2011) and Wakefield (2013). In these textbooks, linear mixed effects models for different types of longitudinal data and methods of estimation are provided. Moreover, penalized spline regression models were introduced by Wahba (1990), Eilers and Marx (1996), Currie and Durban (2002), Durban et al. (2005), Ruppert et al. (2003) and Harrell (2015) to handle

non-linear longitudinal data and smoothing.

2.1.1 Linear mixed effects models

2.1.1.1 Models

Let y_{ij} denote the response variable for the i^{th} individual ($i = 1, .., n$) at the j^{th} occasion ($j = 1, ..., n_i$). Here, n_i is the number of measurements for the i^{th} subject. The vector of the i^{th} individual is denoted by $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})$. The mean at the j^{th} occasion is denoted by $\mu_{ij} = E(y_{ij})$. The covariance between y_{ij} and y_{ik} is denoted by $cov(y_{ij}, y_{ik}) = \sigma_{jk} = E\{(y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})\}$. According to Verbeke and Molenberghs (2000) and Fitzmaurice et al. (2004), the linear mixed effects model can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i\,.$$

Here, $\boldsymbol{X}_i$ is a $(n_i \times p)$ matrix of covariates of fixed effects, $\boldsymbol{Z}_i$ is a $(n_i \times q)$ matrix of covariates of random effects. The columns of the matrix $\boldsymbol{Z}_i$ are a subset of the columns of the matrix $\boldsymbol{X}_i$ $(q \leq p)$. The term $\boldsymbol{X}_i\boldsymbol{\beta}$ is assumed to be shared by all individual. The term $\boldsymbol{Z}_i\boldsymbol{b}_i$ captures the differences between the mean response of the population and individual response trajectories over time. $\boldsymbol{\beta}$ is a $(p \times 1)$ coefficient vector of fixed effects, and $\boldsymbol{b}_i$ is a $(q \times 1)$ vector of random effects.

There are some key assumptions for the linear mixed effects models (Hand and Crowder, 1996; Fitzmaurice et al., 2004). The first assumption is that the vector of random effects, $\boldsymbol{b}_i$, is assumed to have a multivariate normal distribution ($\mathcal{MVN}$) with mean zero and covariance matrix $\boldsymbol{G}$. This means $E(\boldsymbol{b}_i) = 0$ and $cov(\boldsymbol{b}_i) = \boldsymbol{G}$, $i = 1, ..., n$. The second assumption is that the vector of errors, $\boldsymbol{\varepsilon}_i$, is also assumed to have a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{R}_i$. This means $E(\boldsymbol{\varepsilon}_i) = 0$ and $cov(\boldsymbol{\varepsilon}_i) = \boldsymbol{R}_i$, $i = 1, ..., n$.

Based on these assumptions, the conditional expectation of $\boldsymbol{y}_i$ given $\boldsymbol{b}_i$, is $E(\boldsymbol{y}_i|\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ and the conditional covariance of $\boldsymbol{y}_i$, given $\boldsymbol{b}_i$, is $cov(\boldsymbol{y}_i|\boldsymbol{b}_i) = cov(\boldsymbol{\varepsilon}_i) = \boldsymbol{R}_i$. In addition, the population mean of $\boldsymbol{y}_i$ is

$$\begin{aligned} E(\boldsymbol{y}_i) = \mu_i &= E(E(\boldsymbol{y}_i|\boldsymbol{b}_i)) \\ &= E(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i) \\ &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_iE(\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta}\,, \end{aligned}$$

and the covariance of $\boldsymbol{y}_i$, denoted as $\sum_i$, has the form

$$\begin{aligned}\sum{}_i = cov(\boldsymbol{y}_i) &= cov(\boldsymbol{Z}_i\boldsymbol{b}_i) + cov(\boldsymbol{\varepsilon}_i) \\ &= \boldsymbol{Z}_i cov(\boldsymbol{b}_i)\boldsymbol{Z}_i^T + cov(\boldsymbol{\varepsilon}_i) \\ &= \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^T + \boldsymbol{R}_i\,.\end{aligned}$$

2.1.1.2 Parameter estimation

The estimation of the linear mixed-effects models is based on the maximum likelihood (ML) for the fixed effects, the restricted maximum likelihood (REML) for the covariance matrix $\sum_i$ and the best linear unbiased predictor (BLUP) for random effects (Hand and Crowder, 1996; Fitzmaurice et al., 2004; Verbeke and Molenberghs, 2000; Wakefield, 2013). By assuming that the repeated measurements in the longitudinal outcome are independent of each other, the log-likelihood function of the linear mixed effects models has the form

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i; \boldsymbol{\theta})\,,$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \sum_i\right)$ denotes the full parameter vector of the models, and

$$p(\boldsymbol{y}_i; \boldsymbol{\theta}) = (2\pi)^{-n_i/2} |\sum{}_i|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \sum_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}\,.$$

Here $|\boldsymbol{A}|$ denotes the determinant of the matrix $\boldsymbol{A}$. According to Verbeke and Molenberghs (2000) and Fitzmaurice et al. (2004), assuming $\sum_i$ is known, the maximum likelihood estimator of the vector of the fixed effects, $\boldsymbol{\beta}$, has a closed form

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T \sum_i^{-1} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^T \sum_i^{-1} \boldsymbol{y}_i\,. \tag{2.1.1}$$

The estimated covariance matrix of the coefficient vector $\boldsymbol{\beta}$ is

$$\hat{var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T \sum_i^{-1} \boldsymbol{X}_i\right)^{-1}\,.$$

According to Fitzmaurice et al. (2004) and Hand and Crowder (1996), the maximum likelihood estimate of $cov(\boldsymbol{y}_i) = \sum{}_i$ is biased on small samples. Hence, the restricted maximum likelihood method is recommended for estimating $\sum_i$. In particular, if the

coefficient vector, $\boldsymbol{\beta}$, is given, the estimate of $\sum_i$ is obtained by maximizing the slightly modified log-likelihood function having the form

$$\begin{aligned} l(\boldsymbol{G}, \boldsymbol{R}_i) = &- \frac{1}{2}\log\left|\sum_{i=1}^{n} \boldsymbol{X}_i^T \sum\nolimits_i^{-1} \boldsymbol{X}_i\right| - \frac{1}{2}\log\sum_{i=1}^{n}\left|\sum\nolimits_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{n}\left\{(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \sum\nolimits_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}. \end{aligned} \tag{2.1.2}$$

Finally, following Verbeke and Molenberghs (2000) and Fitzmaurice et al. (2004), the estimator of the vector of the random effects using the best linear unbiased predictors (BLUPs) is denoted as $\hat{\boldsymbol{b}}_i$. Based on the assumptions for the linear mixed effects models, we have

$$\begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{\varepsilon}_i \end{bmatrix} \sim \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & 0 \\ 0 & \boldsymbol{R}_i \end{bmatrix}\right).$$

We note that

$$\begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{b}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z}_i & \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{\varepsilon}_i \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{b}_i \end{bmatrix} &\sim \mathcal{MVN}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z}_i & \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{G} & 0 \\ 0 & \boldsymbol{R}_i \end{bmatrix}\begin{bmatrix} \boldsymbol{Z}_i^T & \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix}\right) \\ &\sim \mathcal{MVN}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^T + \boldsymbol{R}_i & \boldsymbol{Z}_i\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{Z}_i^T & \boldsymbol{G} \end{bmatrix}\right). \end{aligned}$$

The BLUP estimator of $\boldsymbol{b}_i$ is $\hat{\boldsymbol{b}}_i$ which has the form

$$\begin{aligned} \hat{\boldsymbol{b}}_i = E(\boldsymbol{b}_i|\boldsymbol{y}_i) &= \boldsymbol{G}\boldsymbol{Z}_i^T(\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^T + \boldsymbol{R}_i)^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\ &= \boldsymbol{G}\boldsymbol{Z}_i'\sum\nolimits_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \end{aligned}$$

where $\sum_i^{-1} = (\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^T + \boldsymbol{R}_i)^{-1}$. The estimate for the variance of the random effects and the error terms has the form

$$\hat{var}(\hat{\theta}_{\boldsymbol{b}_i,\boldsymbol{\epsilon}_i}) = \left\{E\left(-\frac{\partial^2 l(\theta)}{\partial\theta_{\boldsymbol{b}_i,\boldsymbol{\epsilon}_i}^T \partial\theta_{\boldsymbol{b}_i,\boldsymbol{\epsilon}_i}}|_{\theta_{\boldsymbol{b}_i,\boldsymbol{\epsilon}_i}=\hat{\theta}_{\boldsymbol{b}_i,\boldsymbol{\epsilon}_i}}\right)\right\}.$$

According to Fitzmaurice et al. (2004), the BLUP estimator of the random effects has the following properties:

i. $\hat{\boldsymbol{b}}_i$ is a linear function of $\boldsymbol{y}_i$;

ii. $\hat{\boldsymbol{b}}_i$ is unbiased for $\boldsymbol{b}_i$ so that $E(\hat{\boldsymbol{b}}_i - \boldsymbol{b}_i) = 0$;

iii. $var(\hat{\boldsymbol{b}}_i - \boldsymbol{b}_i)$ is no larger than the $var(\tilde{\boldsymbol{b}}_i - \boldsymbol{b}_i)$ where $\tilde{\boldsymbol{b}}_i$ is any other linear and unbiased predictor.

2.1.2 Penalized spline longitudinal models

When subjects show non-linear longitudinal trajectories, it is necessary to consider flexible non-linear regressions. Penalized spline regression models are considered as extensions of linear regression models to handle such non-linear longitudinal relationships (Ruppert et al., 2003; Currie and Durban, 2002; Durban et al., 2005; Wahba, 1990). These models have become effective ways of handling highly non-linear trajectories, especially when a large number of knots are inserted into the model.

Recall that y_{ij} denotes the longitudinal response for the i^{th} subject , $i = 1, ..., n$ which is measured at time point t_{ij}, $j = 1, ..., n_i$. According to Ruppert et al. (2009), the general spline model of degree p has the form

$$y_{ij} = f(t_{ij}) = \beta_0 + \beta_1 t_{ij} + ... + \beta_p t_{ij}^p + \sum_{k=1}^{K} u_{pk}(t_{ij} - \mathcal{K}_k)_+^p + \varepsilon(t_{ij}) \,, \tag{2.1.3}$$

where the set $\left\{1, t_{ij}, ..., t_{ij}^p, (t_{ij} - \mathcal{K}_1)_+^p, ..., (t_{ij} - \mathcal{K}_K)_+^p\right\}$ is known as the truncated power basis of degree p, and the function $(.)_+$ is defined by $(x)_+ = max(0, x)$, for all real x. The vector $\boldsymbol{\beta}^T = (\beta_0, ..., \beta_p, u_{p1}, ..., u_{pK})$ is the $((p + K + 1) \times 1)$ row vector of coefficients. Moreover, $\mathcal{K}_1, ..., \mathcal{K}_K$ are fitted K knots. The assumption for the measurement error is normal distribution $\varepsilon(t_{ij}) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. Now, we write the model (2.1.3) in matrix notation as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,, \tag{2.1.4}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{nn_n} \end{bmatrix}, \; \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix},$$

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & \cdots & t_{i1}^p & (t_{i1} - \mathcal{K}_1)_+^p & \cdots & (t_{i1} - \mathcal{K}_K)_+^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & \cdots & t_{in_i}^p & (t_{in_i} - \mathcal{K}_1)_+^p & \cdots & (t_{in_i} - \mathcal{K}_K)_+^p \end{bmatrix}.$$

Two problems need to be carefully considered in Model (2.1.3). The first is that this model may cause roughness of the fit. If there is a large set of knots inserted into the model, the fitted function can have small random fluctuations. The second is that the nonparametric function $f(.)$ is for the population mean and does not depend on the individual. Therefore, the model in (2.1.3) needs to be extended to model subject specific curves.

The roughness of the fit is due to the existence of too many knots in the model, which can lead to an over-fitted function (Good and Gaskins, 1971). To solve this problem, Ruppert et al. (2003) suggested that all the knots be retained, but the coefficients of the knots be constrained. This will restrict the influence of the variables $(x - \mathcal{K}_k)_+^p$ and will lead to smoother spline functions. Hence, the estimation problem is to choose $\boldsymbol{\beta}$ to minimize $\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2$ with constraints on the u_{pk}.

Alternatively, suppose we define $\boldsymbol{D}$ to be the $(K + p + 1) \times (K + p + 1)$ diagonal matrix with the form

$$\boldsymbol{D} = \begin{bmatrix} 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1_K \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{p+1\times p+1} & \mathbf{0}_{p+1\times K} \\ \mathbf{0}_{K\times p+1} & \mathbf{1}_{K\times K} \end{bmatrix}.$$

Following this, the problem is to choose $\boldsymbol{\beta}$ to minimize $\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2$ subject to $\boldsymbol{\beta}^T D \boldsymbol{\beta} \leq C$. By using a Lagrange multiplier argument, this is equivalent to choosing $\boldsymbol{\beta}$ to minimize

$$\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 + \lambda \boldsymbol{\beta}^T D \boldsymbol{\beta}\,, \tag{2.1.5}$$

for a suitable number $\lambda \geq 0$. The term $\lambda \boldsymbol{\beta}^T D \boldsymbol{\beta}$ is called a roughness penalty, and λ is known as the smoothing parameter. The amount of smoothing is controlled by λ. Ordinary least squares corresponds to $\lambda = 0$, where the u_{pk} are unrestricted. When λ is taken as a positive finite value, this leads to smaller estimates of the u_{pk} and the effects of $(\boldsymbol{x} - \mathcal{K}_k)_+^p$ are then less. When we take λ to be very large, the effects of the knots diminishes and the model becomes the least squares line.

To determine the smoothing parameter λ, Ruppert et al. (2003) and Durban et al. (2005) considered penalized splines as mixed models. In particular, we have the form of the

general spline models as in (2.1.3). First we define $\boldsymbol{\beta}^T = [\beta_0, ..., \beta_p]$ as a $((p+1) \times 1)$ row vector of fixed effects, and $\boldsymbol{b}^T = [u_{p1}, ..., u_{pK}]$ as a $(K \times 1)$ row vector of random effects. The mixed effects regression model is then given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}\,, \tag{2.1.6}$$

where

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix}, \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 \\ \vdots \\ \boldsymbol{Z}_n \end{bmatrix},$$

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & \cdots & t_{i1}^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & \cdots & t_{in_i}^p \end{bmatrix}, \boldsymbol{Z}_i = \begin{bmatrix} (t_{i1} - \mathcal{K}_1)_+^p & \cdots & (t_{i1} - \mathcal{K}_K)_+^p \\ \vdots & \vdots & \vdots \\ (t_{in_i} - \mathcal{K}_1)_+^p & \cdots & (t_{in_i} - \mathcal{K}_K)_+^p \end{bmatrix}.$$

The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are respectively designed matrices of fixed effects covariates and random effects covariates. We assume that $\boldsymbol{y} \mid \boldsymbol{b} \sim \mathcal{MVN}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \sigma_\varepsilon^2\mathbf{I})$ and $\boldsymbol{b} \sim \mathcal{MVN}(0, \sigma_u^2\mathbf{I})$.

Under these assumptions, the log-likelihood function of the model has the form

$$\begin{aligned} \log\{p(y, b; \theta)\} &= \log\{p(y \mid b; \theta)p(b; \theta)\} \\ &= \log\left[\frac{1}{\sqrt{2\pi}\sigma_\varepsilon} exp\left\{\frac{\parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b} \parallel^2}{\sigma_\varepsilon^2}\right\} \frac{1}{\sqrt{2\pi}\sigma_u} exp\left\{\frac{\parallel \boldsymbol{b} \parallel^2}{\sigma_u^2}\right\}\right]. \end{aligned} \tag{2.1.7}$$

Therefore, for the model in (2.1.6), the main aim is to obtain the estimate for the unknowns $\boldsymbol{\beta}$ and $\boldsymbol{b}$ that minimizes

$$\frac{1}{\sigma_\varepsilon^2} \parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b} \parallel^2 + \frac{1}{\sigma_u^2} \parallel \boldsymbol{b} \parallel^2 . \tag{2.1.8}$$

By comparing equations (2.1.5) and (2.1.8), the smoothing parameter is obtained as $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_u^2}$.

To specify the individual curves, Ruppert et al. (2003) and Durban et al. (2005) presented flexible models for which each individual has its own function. The penalized spline model for subject-specific curves has the form

$$\begin{aligned} y_{ij} &= f(t_{ij}) + g_i(t_{ij}) + \varepsilon(t_{ij}),\ \varepsilon(t_{ij}) \sim \mathcal{N}(0, \sigma_\varepsilon^2), \\ g_i(t_{ij}) &= b_{i0} + b_{i1}t_{ij} + b_{i2}t_{ij}^2 + ... + b_{ip}t_{ij}^p + \sum_{k=1}^{K} v_{ipk}(t_{ij} - \mathcal{K}_k)_+^p\,, \end{aligned} \tag{2.1.9}$$

where the $f(.)$ function is as in (2.1.3). This model can be described in the mixed model framework as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}\,,$$

where

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix}, \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{X}_1 & 0 & \dots & 0 & \boldsymbol{Z}_1 & 0 & \dots & 0 \\ \boldsymbol{Z}_2 & 0 & \boldsymbol{X}_2 & \dots & 0 & 0 & \boldsymbol{Z}_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{Z}_n & 0 & 0 & \dots & \boldsymbol{X}_n & 0 & 0 & \dots & \boldsymbol{Z}_n \end{bmatrix},$$

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & \cdots & t_{i1}^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & \cdots & t_{in_i}^p \end{bmatrix}, \boldsymbol{Z}_i = \begin{bmatrix} (t_{i1} - \mathcal{K}_1)_+^p & \cdots & (t_{i1} - \mathcal{K}_K)_+^p \\ \vdots & \vdots & \vdots \\ (t_{in_i} - \mathcal{K}_1)_+^p & \cdots & (t_{in_i} - \mathcal{K}_K)_+^p \end{bmatrix}.$$

$$\boldsymbol{b}^T = (u_{p1}, ..., u_{pK}, b_{10}, ..., b_{1p}, ..., b_{n0}, ..., b_{np}, v_{1p1}, ..., v_{1pK}, ..., v_{npK}, ..., v_{npK})\,,$$

$$\boldsymbol{\beta}^T = (\beta_0, ..., \beta_p)\,.$$

Ruppert et al. (2003) assumed that $(b_{i0}, ..., b_{ip})^T \sim \mathcal{MVN}(0, \sum)$ and v_{ipk} follows an univariate normal distribution ($\mathcal{UVN}$), $v_{ipk} \sim \mathcal{UVN}(0, \sigma_v^2)$. Then, the covariance matrix of the random effects is

$$G = cov(\boldsymbol{b}) = \begin{bmatrix} \sigma_u^2 \boldsymbol{I} & 0 & 0 \\ 0 & \underset{1 \le i \le m}{block} \sum & 0 \\ 0 & 0 & \sigma_v^2 \boldsymbol{I} \end{bmatrix}.$$

2.2 Survival analysis of event time data

Recently, survival analysis has been developed extensively in the literature and has been widely used especially in clinical and epidemiological studies. These studies aim to analyze the time until a specified event of interest happens. Cox (1972, 1975), Cox and Hinkley (1979) and Cox and Oakes (1984) introduced a very popular Cox model for survival data. These models assume that time independent covariates have an effect on the hazard function for an event.

Along this line, Kalbfleisch and Prentice (2002); Hougaard (2000); Klein and Moeschberger (2005) provided a general theory for event time data with the survival distributions and

basic statistical tools for their analysis. Andersen et al. (1993) and Aalen et al. (2008) presented a more theoretical analysis for the Cox model using martingales and counting processes. Another trend for survival analysis focuses on statistical modelling and estimating techniques (Therneau and Grambsch, 2000; Ibrahim et al., 2005; Rizopoulos, 2012, 2010, 2014). They proposed more flexible joint models for different types of longitudinal data and a censoring mechanism as well as estimation methods.

In this section, we present the basic functions and the special features of survival data (Kalbfleisch and Prentice, 2002; Andersen et al., 1993) in Sections 2.2.1 and 2.2.2. In addition, we review the famous Cox model for time independent covariates and extended Cox models for time dependent covariates (Cox, 1972, 1975; Cox and Hinkley, 1979; Cox and Oakes, 1984) in Section 2.2.3.

2.2.1 Basic functions of survival data

Let T denote the random variable of failure times, which is assumed continuous. The three equivalent functions that are usually used to define the distribution function of survival time T are: the survival function $S(t)$, the probability density function $f(t)$ and the hazard function $h(t)$. According to Cox and Oakes (1984) and Aalen et al. (2008), the definition of the survival function is

$$\begin{aligned} S(t) &= \Pr(\text{an individual survives longer than t}) \\ &= \Pr(T > t) = \int_t^\infty f(s)ds\,. \end{aligned}$$

Let $F(t)$ be the cumulative distribution function for survival time T. Then

$$S(t) = 1 - F(t)\,.$$

In addition, if the hazard function is defined as

$$h(t) = \lim_{\triangle t \to 0} \frac{Pr(t \le T < t + \triangle t | T \ge t)}{\triangle t}, \quad t > 0\,,$$

the relationship between the survival function $S(t)$, the probability density function $f(t)$ and the hazard function $h(t)$ can be written as

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} \\ &= -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log S(t)\,, \end{aligned}$$

where $S'(t)$ is the first derivative of the survival function $S(t)$. The cumulative hazard function $H(t)$ is

$$H(t) = \int_0^t h(x)dx = -\log S(t)\,.$$

Hence, the survival function can be written in terms of the cumulative hazard function as

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(x)dx\right\}\,.$$

2.2.2 Exogenous and endogenous covariates

When survival function $S(t)$ is assumed to have a specific parametric form associating with a longitudinal submodel, estimations for parameters of interest are usually based on the likelihood function (Rizopoulos, 2012). In the maximum likelihood method, there are different treatments for different types of covariates in the longitudinal submodel. Here, we present the two different categories of time dependent covariates and the estimation techniques for these covariates will be introduced in the following sections.

We let the time-dependent covariate for the i^{th} subject at time t be denoted by $y_i(t)$. We let $\mathcal{Y}_i(t) = \{y_i(s), 0 \leq s < t\}$ denote the covariate history of the i^{th} subject up to time t. According to Kalbfleisch and Prentice (2002), the exogenous covariates are the covariates satisfying the condition:

$$\Pr\left(s \leq T_i < s + ds | T_i \geq s, \mathcal{Y}_i(s)\right) = \Pr\left(s \leq T_i < s + ds | T_i \geq s, \mathcal{Y}_i(t)\right)\,, \tag{2.2.1}$$

for all s, t such that $0 < s \leq t$, and $ds \to 0$. An equivalent definition is

$$\Pr\left(\mathcal{Y}_i(t) | \mathcal{Y}_i(s), T_i \geq s\right) = \Pr\left(\mathcal{Y}_i(t) | \mathcal{Y}_i(s), T_i = s\right)\,,\ s \leq t\,. \tag{2.2.2}$$

On the other hand, endogenous time-varying covariates are the ones that do not satisfy the condition in (2.2.1). In particular,

$$\Pr\left(\mathcal{Y}_i(t) | \mathcal{Y}_i(s), T_i \geq s\right) \neq \Pr\left(\mathcal{Y}_i(t) | \mathcal{Y}_i(s), T_i = s\right)\,,\ s \leq t\,.$$

Based on the definitions in (2.2.1) and (2.2.2), the future path of exogenous covariates up to time $t \geq s$ does not affect the hazard rate at time s. Its value at any time t is predicted

before t. Moreover, under the conditions (2.2.1) and (2.2.2), one can define the survival function conditional on the covariate path

$$\begin{aligned} S_i\left(t|\mathcal{Y}_i(t)\right) &= \Pr\left(T_i > t|\mathcal{Y}_i(t)\right) \\ &= \exp\left(-\int_0^t h_i\left(s|\mathcal{Y}_i(s)\right) ds\right) . \end{aligned} \tag{2.2.3}$$

According to Kalbfleisch and Prentice (2002) and Rizopoulos (2012), there are some important features of endogenous covariates which are different from exogenous covariates. Firstly, the future path of endogenous covariates is not predictable. The second is that its value at time point t shows the survival of the subject at this time. In particular, when failure is defined as the death of the subject,

$$S_i\left(t|\mathcal{Y}_i(t)\right) = Pr\left(T_i^* > t|\mathcal{Y}_i(t)\right) = 1 , \tag{2.2.4}$$

if $y_i(t-ds)$ is given with $ds \to 0$. Due to this feature, the log-likelihood based on $f(t)$ and $S(t)$ is not suitable for endogenous covariates. Another feature of endogenous covariates is that they contain measurement errors.

2.2.3 The Cox and extended Cox models

The Cox and extended Cox models are the models which were proposed to link between exogenous covariates and survival time using proportional hazards models (Cox, 1972). The Cox model handles independent time covariates whereas the extended Cox model handles external time-dependent covariates. For both models, the partial likelihood method is usually implemented to estimate the parameters in the models.

Suppose that there are n subjects in the longitudinal data and survival data. The observed failure time for the i^{th} subject is denoted as $T_i = \min(T_i^*, C_i)$. Here, T_i^* is the true survival time and C_i denotes the censoring time for the i^{th} subject $(i = 1, ..., n)$. An event indicator is also defined as $\delta_i = I(T_i^* \leq C_i)$ in survival data. The longitudinal data consists of the measurements of the subjects.

The proportional hazards model proposed by Cox (1972) has the form

$$\begin{aligned} h(t \mid \boldsymbol{z}) &= h_0(t) \exp(z_1\beta_1 + ... + z_p\beta_p) \\ &= h_0(t) \exp(\boldsymbol{z}^T\boldsymbol{\beta}) . \end{aligned} \tag{2.2.5}$$

Here, $h_0(t)$ is the hazard at baseline, $\boldsymbol{z}$ is a $p \times 1$ vector of covariates and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. Obviously,

$$h(t|\boldsymbol{z} = 0) = h_0(t)\,.$$

$h_0(t)$ can be interpreted as the hazard function for the population of subjects with $\boldsymbol{z} = 0$.

According to Cox (1972, 1975), the partial likelihood function, PL(.), can be written as

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(z_i^T \boldsymbol{\beta})}{\sum_{l=1}^{n} \exp(z_l^T \boldsymbol{\beta}) Y_l(t_i)} \right]^{\delta_i} .$$

Here, $t_1, ..., t_n$ define the distinct death times and $Y_i(t)$ denotes the indicator for whether or not the i^{th} individual is at risk at time t. It can be seen that the value of the covariates are only required at the event times, and these covariates are independent of time in the Cox model. Therefore, the model cannot handle the time dependent covariates.

The Cox model was then extended to handle external time-dependent covariates using a counting process as in Cox and Hinkley (1979); Cox and Oakes (1984); Andersen et al. (1993). In the counting process notation, the event process for the i^{th} subject is written as $\{N_i(t), Y_i(t)\}$, where $N_i(t)$ denotes the number of events for subject i by time t, and $Y_i(t)$ denotes the indicator for whether or not the i^{th} individual is at risk at time t. The extended Cox model is written as

$$h_i(t \mid \mathcal{Y}_i(t), \boldsymbol{w}_i) = h_0(t) Y_i(t) \exp\left\{ \boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha y_i(t) \right\} . \tag{2.2.6}$$

Here, $h_0(t)$ is the hazard at baseline, and $\boldsymbol{w}_i$ is a vector of baseline covariates. Furthermore, $\mathcal{Y}_i(t) = \{m_i(s),\, 0 \leq s < t\}$ denotes the history of the true unobserved longitudinal process up to time t.

Estimation of $\boldsymbol{\gamma}$ and α in (2.2.6) is based on the partial likelihood function (Kalbfleisch and Prentice, 2002) that can be written as

$$PL(\boldsymbol{\gamma}, \alpha) = \prod_{i=1}^{n} \prod_{\{\text{all grid point u}\}} \left[\frac{Y_i(u) \exp\left\{ \boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha y_i(u) \right\}}{\sum_{l=1}^{n} Y_l(u) \exp\left\{ \boldsymbol{\gamma}^T \boldsymbol{w}_l + \alpha y_l(u) \right\}} \right]^{dN_i(u)} .$$

Here, the time axis is broken into a grid of points. We can choose the grid points dense enough so that at most one death can occur within any interval. Equivalently, the partial

log-likelihood function can be rewritten as

$$\log(PL(\boldsymbol{\gamma},\alpha)) = \sum_{i=1}^{n}\int_{0}^{\infty}\Bigg\{\log\left\{Y_i(t)\exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_i+\alpha y_i(t)\right\}\right\} - \log\left[\sum_{l=1}^{n}Y_l(t)\exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_l+\alpha y_l(t)\right\}\right]\Bigg\}dN_i(t).$$

2.3 Standard joint models for longitudinal and time-to-event data

2.3.1 Standard joint models

Longitudinal data and survival data are usually recorded together in practice. In many biomarker research and clinical studies, endogenous time-dependent covariates have been recorded along with the survival time. However, the extended Cox models are only suitable to handle exogenous time-dependent covariates. A number of statisticians have recently paid attention to the association between endogenous time-dependent covariates and survival data. The joint modelling framework was introduced in order to handle this primary interest. This modelling framework was proposed by Faucett and Thomas (1996); Tsiatis and Davidian (2001); Henderson et al. (2000); Tsiatis et al. (1995); Rizopoulos (2012). They not only develop the statistical modelling but also show different methods for parameter estimation. Faucett and Thomas (1996) and Rizopoulos (2014) used a Bayesian approach whereas Tsiatis et al. (1995), Tsiatis and Davidian (2001) and Rizopoulos (2012) proposed the frequentist approach.

In this section, we review the standard joint models for longitudinal and time-to-event data. This review includes the two submodels within the joint models: the survival and longitudinal submodels. Following this, parameter estimation using a classical approach is then reviewed. In particular, we provide a full likelihood approach for estimating parameters in the joint models (Rizopoulos, 2012, 2010, 2011; Henderson et al., 2000).

2.3.1.1 The survival submodel

Recall the notions presented in Section 2.2.3. T_i^* denotes the true event time for the i^{th} subject, T_i is the observed event time, which is the minimum of the censoring time C_i, and $T_i^*$and $\delta_i = I(T_i^* \leq C_i)$ is the event indicator. Tsiatis and Davidian (2001) and Rizopoulos (2012) introduced the new term $m_i(t)$, which is the true unobserved longitudinal value of the i^{th} subject at time t. Then they defined the proportional hazards model to link the hazard rate and $m_i(t)$. The risk model has the form

$$\begin{aligned} h_i(t|\mathcal{M}_i(t), \boldsymbol{w}_i) &= \lim_{dt \to 0} Pr\left\{t \leq T_i^* < t + dt | \mathcal{M}_i(t), \boldsymbol{w}_i\right\} / dt \\ &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}, \quad t > 0\,, \end{aligned} \tag{2.3.1}$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ denotes the history of $m_i(t)$ up to time point t, $h_0(.)$ denotes the baseline hazard function, and $\boldsymbol{w}_i$ is the vector of baseline covariates. The parameters $\boldsymbol{\gamma}$ and α quantify the effect of baseline covariates and the longitudinal outcome to the risk of an event. Using the relation between the hazard function, the survival function and the cumulative hazard function, we have

$$\begin{aligned} S_i(t|\mathcal{M}_i(t), \boldsymbol{w}_i) &= Pr(T_i^* > t | \mathcal{M}_i(t), \boldsymbol{w}_i) \\ &= \exp\left(-\int_0^t h_0(s) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s)\right\} ds\right). \end{aligned} \tag{2.3.2}$$

We need to discuss the form for the baseline hazard function, $h_0(t)$, in order to have the complete form of the risk model in (2.3.1). In the standard survival analysis, $h_0(t)$ has a completely unspecified form (Cox and Oakes, 1984). However, within the joint modelling framework, the form of $h_0(t)$ needs to be specified in order to calculate the standard errors of parameter estimates.

There are two simple options that usually work quite satisfactorily in practice for defining $h_0(.)$. The first option is to choose a standard distribution for the hazard rate at the baseline. Typical distributions used for $h_0(t)$ are the exponential distribution, the Gompertz distribution, and the Weibull distribution Cox and Oakes (1984); Crowther and Lambert (2013). The second option is to use a semiparametric approach for the hazard rate at the baseline. Among these are the piecewise-constant and regression splines approaches Rizopoulos (2012); Ibrahim et al. (2005).

2.3.1.2 The longitudinal submodel

Let $y_i(t)$ denote the observed longitudinal value for the i^{th} subject at time t. All measurements for the i^{th} subject are $\{y_i(t_{ij}),\ j = 1, ..., n_i\}$. According to Tsiatis et al. (1995); Tsiatis and Davidian (2001); Rizopoulos (2010), the association between $y_i(t)$ and $m_i(t)$ is defined through the longitudinal submodel as

$$\begin{cases} y_i(t) & = m_i(t) + \varepsilon_i(t),\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ m_i(t) & = \boldsymbol{X}_i^T(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^T(t)\boldsymbol{b}_i \\ \boldsymbol{b_i} & \sim \mathcal{MVN}(0, \boldsymbol{D})\,, \end{cases} \tag{2.3.3}$$

where $\boldsymbol{X}_i(t)$ is a designed matrix of covariates of fixed effects and $\boldsymbol{Z}_i(t)$ is a designed matrix of covariates of random effects. In addition, $\boldsymbol{\beta}$ is a coefficient vector of fixed effects and $\boldsymbol{b}_i$ is a vector of random effects. Moreover, we assume that the error term, $\varepsilon_i(t)$, follows a normal distribution with mean 0 and variance σ_ε^2. The measurement error is independent of the random effects $\boldsymbol{b_i}$ which follows the multivariate normal distribution with mean 0 and covariance matrix $\boldsymbol{D}$.

2.3.2 Frequentist inference

In frequentist approaches, the Cox and extended Cox methods as presented in Section 2.2.3 are some of the simplest methods for estimating paramaters in the joint models. In these methods, the estimation for parameters is based on maximizing the partial likelihood function. However, there are assumptions for these models which cause bias and are unrealisitic (Sweeting and Thompson, 2011; Rizopoulos, 2012). The time-dependent covariates are assumed to be constant in the interval between the visiting times. Time-dependent covariates are predicted processes and measured without error. In this section, we present two more classical approaches for joint models, namely an ordinary two-stage approach and a full likelihood approach.

2.3.2.1 An ordinary two-stage approach

An ordinary two-stage approach has been investigated in Tsiatis et al. (1995); Tsiatis and Davidian (2001); Bycott and Taylor (1998). In this approach, there are two stages for

estimating parameters in the standard joint models. In the first stage, they used the linear mixed effects model to fit only the longitudinal process. The maximum likelihood estimation and the BLUPs are used to estimate the longitudinal coefficients and random effects. Then, in the second stage, the longitudinal fitted values are considered as covariates in the survival submodel. The partial likelihooad approach is applied to estimate the survival cofficients and the hazard rate at baseline.

In the first stage, the fitted longitudinal model has a form

$$\hat{m}_i(t) = \boldsymbol{X}_i^T(t)\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i^T(t)\hat{\boldsymbol{b}}_i\,.$$

In the second stage, the partial likelihood has a form

$$\log(PL(\boldsymbol{\gamma},\alpha)) = \sum_{i=1}^{n}\int_0^{\infty}\Bigg\{\log\left\{R_i(t)exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_i + \alpha\hat{m}_i(t)\right\}\right\} - \log\left[\sum_{l=1}^{n}R_l(t)exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_l + \alpha\hat{m}_l(t)\right\}\right]\Bigg\}dN_i(t)\,.$$

Here, $R_i(t) = 1$ if the i^{th} subject is at risk at time t. Otherwise, $R_i(t) = 0$.

Since the estimated longitudinal process, $\hat{m}_i(t)$, is continuous throughout time, the grid points can be choosen as fine as required. Therefore, the assumption of constant longitudinal measurements between the visiting times is weakened. The another obvious advantage of using a two-stage approach is its quick implementation. Tsiatis et al. (1995) used standard linear mixed effects and survival software for the first stage and the second stage respectively. However, this approach has problems when subjects suffer informative drop-out. Moreover, the method strongly depends on the normality assumptions for random effects and error terms in the first stage. The drawbacks of this approach were discussed in detail by Tsiatis and Davidian (2001); Sweeting and Thompson (2011).

2.3.2.2 A full likelihood approach

To define the joint likelihood function for the standard joint models as in Section 2.3.1, some key assumptions for random effects and the visiting process have been proposed by Rizopoulos (2012). One assumption is that the vector of time-dependent random effects

$\boldsymbol{b_i}$ has an effect on both the longitudinal and survival processes. Formally,

$$\begin{aligned} p(T_i, \delta_i, \boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) &= p(T_i, \delta_i|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) \\ p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) &= \prod_j p\left\{\boldsymbol{y_i}(t_{ij})|\boldsymbol{b_i}; \boldsymbol{\theta}\right\}, \end{aligned} \tag{2.3.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_t^T, \boldsymbol{\theta}_y^T, \boldsymbol{\theta}_b^T)$ denotes the full parameter vector, with $\boldsymbol{\theta}_t = (\gamma^T, \alpha, \theta_{h_0}^T)^T$ denoting the parameters for the survival outcome, $\boldsymbol{\theta}_y$ the parameters for longitudinal outcomes, and $\boldsymbol{\theta}_b$ the variance matrix of random effects. In addition, the censoring mechanism and the visiting process are assumed to be independent of the true event times and future longitudinal measurements.

Under these assumptions, the log-likelihood function of the joint models has the form

$$\begin{aligned} \log p(T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}) &= \log \int p(T_i, \delta_i, \boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) d\boldsymbol{b_i} \\ &= \log \int p(T_i, \delta_i|\boldsymbol{b_i}; \boldsymbol{\theta}) \left[\prod_j p\left\{\boldsymbol{y_i}(t_{ij})|\boldsymbol{b_i}; \boldsymbol{\theta}\right\}\right] p(\boldsymbol{b_i}; \boldsymbol{\theta}_b) d\boldsymbol{b_i}. \end{aligned} \tag{2.3.5}$$

Here, the conditional density for the survival part has the form

$$\begin{aligned} p(T_i, \delta_i|\boldsymbol{b_i}; \boldsymbol{\theta}) &= h_i(T_i|\mathcal{M}(T_i); \boldsymbol{\theta}_t, \beta)^{\delta_i} S_i(T_i|\mathcal{M}(T_i); \boldsymbol{\theta}_t, \beta) \\ &= \left[h_0(T_i) exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(T_i)\right\}\right]^{\delta_i} \exp\left(-\int_0^{T_i} h_0(s) exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s)\right\} ds\right). \end{aligned} \tag{2.3.6}$$

On the other hand, the density for the longitudinal part with the random effects is given by

$$\begin{aligned} p(\boldsymbol{y_i}|\boldsymbol{b}_i; \boldsymbol{\theta}) &= \prod_j p\left\{y_i(t_{ij})|\boldsymbol{b_i}; \boldsymbol{\theta}\right\} p(\boldsymbol{b_i}; \boldsymbol{\theta}_b) \\ &= (2\pi\sigma^2)^{-\frac{n_i}{2}} \exp(- \parallel \boldsymbol{y_i} - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b_i} \parallel^2 / 2\sigma_\varepsilon^2) \\ &\quad \times (2\pi)^{-\frac{q_b}{2}} det(D)^{-1/2} \exp(-\boldsymbol{b}_i^T D^{-1} \boldsymbol{b_i}/2), \end{aligned} \tag{2.3.7}$$

where q_b denotes the dimension of the random effects $\boldsymbol{b_i}$.

The observed data score vector for the joint models can be written as

$$\begin{aligned}
S(\boldsymbol{\theta}) &= \sum_i \frac{\partial}{\partial \boldsymbol{\theta}^T} \log \int p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i \\
&= \sum_i \frac{1}{p(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}^T} \int p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i \\
&= \sum_i \frac{1}{p(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta})} \int \frac{\partial}{\partial \boldsymbol{\theta}^T} p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i \\
&= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}^T} \log \{p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\} \\
&\qquad \times \frac{p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)}{p(T_i, \delta_i, \boldsymbol{y}_i; \theta)} d\boldsymbol{b}_i \\
&= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}^T} \log \{p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\} \, p(\boldsymbol{b}_i \mid T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i \, .
\end{aligned} \tag{2.3.8}$$

To estimate the parameters in model (2.3.1), Rizopoulos (2010, 2011) used the EM algorithm. In particular, to derive the maximum likelihood estimates in (2.3.1), the algorithm obtained the parameter estimates $\hat{\boldsymbol{\theta}}$ which maximize instead the expected value of the complete data log-likelihood at the i^{ih} iteration of

$$\begin{aligned}
Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(it)}) &= \sum_i \int \log \left(p(T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b}_i; \boldsymbol{\theta})\right) . p(\boldsymbol{b}_i | T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i \\
&= \sum_i \int \left(\log p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{y_i} | \boldsymbol{b_i}; \boldsymbol{\theta}) + \log p(\boldsymbol{b}_i; \boldsymbol{\theta})\right) p(\boldsymbol{b}_i | T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i \, .
\end{aligned} \tag{2.3.9}$$

To support this approach, Rizopoulos (2010) introduced the popular R package JM in which the standard joint models are used to fit for the longitudinal and time-to-event data. This full likelihood approach provides better results in the frequentist approaches for joint models (Sweeting and Thompson, 2011; Gould et al., 2014). For linear longitudinal simulated data, Sweeting and Thompson (2011) showed that this approach provides unbiased results compared to the Cox model approach and ordinary two-stage approach.

However, there are some disadvantages which these joint models have to deal with. Computational complexity is one of the problems. The multi-integrals in (2.3.9) do not usually have closed form solutions. The computational burden will increase dramatically when the dimension of the random effects is large Rizopoulos (2011). Because of this, the maximum dimension of random effects in the JM package is set to four (Rizopoulos, 2010). Another problem is that flexible joint models are required when subject-specific trajectories show

non-linear curves over time (Gould et al., 2014). Furthermore, overfitting problems also need to be considered in the joint modelling framework.

2.4 Bayesian inference

There are some advantages in applying a Bayesian approach compared to a frequentist approach for the joint models. In the frequentist approach for the joint models presented in Section 2.3, the parameter estimations are based on the joint likelihood function. This approach has to handle multiple integrals with respect to the random effects appearing in the two submodels. This can lead to computational complexity as noted previously (Rizopoulos, 2012).

In a Bayesian approach, the asymptotic approximations for the integral solution are not needed. Instead, the parameters are estimated using a set of Markov chain Monte Carlo (MCMC) algorithms to approximate the joint posterior distribution Ibrahim et al. (2005); Huang (2009). A Bayesian approach provides a more straightforward way to estimate parameters in terms of computational implementation. In addition to this, the parameters in the joint models are treated as unknown constants in the frequentist approach. A Bayesian approach considers all unknown quantities in the joint model as random variables (Gelman et al., 1995). Therefore, the joint posterior distribution for parameters is sampled to quantify uncertainties using a Bayesian approach. Another advantage of a Bayesian approach is that historical information of subjects can be easily added to the inference procedure(Gould et al., 2014).

In this section, we first review the Bayes' rule following Geman and Geman (1984), Gelman et al. (1995), Robert and Casella (2004), Brooks et al. (2011), Wakefield (2013) and Rizopoulos (2014). Following the work of these authors, the prior distribution, the posterior distribution and the proposal distribution for joint models are also presented. Finally, the MCMC algorithms will be introduced based on Metropolis et al. (1953), Hastings (1970), Cox and Hinkley (1979).

2.4.1 Bayes' rule

Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ denote the unknown parameter vector of the joint model and let $\boldsymbol{y} = (y_1, ..., y_n)^T$ be the vector of the observed data. According to Cox and Hinkley (1979); Robert and Casella (2004); Gelman et al. (1995); Wakefield (2013); Brooks et al. (2011), Bayes' rule is presented as in the following equation

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}\,, \tag{2.4.1}$$

where the normalizing constant is

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d(\boldsymbol{\theta})\,.$$

Here, $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the joint posterior probability distribution of $\boldsymbol{\theta}$ given the observed data $\boldsymbol{y}$, the joint prior distribution is $p(\boldsymbol{\theta})$, and the joint likelihood function is $p(\boldsymbol{y}|\boldsymbol{\theta})$. By ignoring the normalizing constant, equation (2.4.1) can be rewritten as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,. \tag{2.4.2}$$

Equivalently,

$$posterior \propto likelihood \times prior\,. \tag{2.4.3}$$

To make inferences through a Bayesian approach, there are three basic steps that need to be implemented (Gelman et al., 1995). Setting up a full probability joint model is the first step. Here, the joint model from (2.3.1) and (2.3.3) is chosen. Calculating and interpreting the joint posterior is the second step. The third step is checking the fit of the model.

In the second step, to calculate the joint posterior distribution, we first obtain the joint likelihood function and specify the joint prior distributions for parameters in the model. Then, the joint posterior distribution is derived using (2.4.3). In this step, when all conditional posterior distributions derived from a joint posterior distribution have closed forms, we can then sample parameters directly. However, when some or all of the conditional posterior distributions have non-standard forms, we need to implement algorithms to sample parameters having a conditional posterior as a stationary distribution. Before detailing the algorithms, the joint likelihood function, the prior distribution and the joint posterior distribution will be derived in the following sections.

2.4.2 The posterior distributions for the joint models

According to Rizopoulos (2014), the joint likelihood function of the joint model in (2.3.1) and (2.3.3) has the form

$$\begin{aligned} p(T, \delta, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b}) &= \prod_{i=1}^{n} p(T_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{n_i} p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}) p(y_{ij}|\boldsymbol{b}_i; \boldsymbol{\theta}) \,. \end{aligned}$$

The joint posterior distribution is defined as

$$p(\boldsymbol{\theta}, \boldsymbol{b}|T, \delta, \boldsymbol{y}) \propto \prod_{i=1}^{n} \prod_{j=1}^{n_i} p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}) p(y_{ij}|\boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{\theta}) \,,$$

where $p(\boldsymbol{\theta})$ is the joint prior distribution for the parameter vector in the joint model. Specification of the prior distribution is important for defining the conditional posterior distribution.

In general, there are three types of prior distribution for which we can choose for the parameters (Gelman et al., 1995; Wakefield, 2013; Robert and Casella, 2004):

(i) Conjugate prior

When the prior information about the parameters in the model is limited, the prior distribution for a parameter is chosen from a parametric family such that the conditional posterior distribution also belongs to this family. These families are called conjugate prior families. The main reason for choosing conjugate priors is because they usually lead to the standard form for the conditional posterior distributions.

(ii) Non-informative prior

When the information about the parameters in the models is not available, the non-informative priors are defined by a Bayesian approach. These prior distributions can be chosen from either proper or improper distributions which contain no special interest for the parameters.

Reference priors, proposed by **?**, provide another non-informative priors choice. A reference prior $p(.)$ is the distribution which maximizes the expected Kullback-Leibler information, where the expected Kullback-Leibler information has the form

$$K(f,g) = \int \log\left[\frac{p(\boldsymbol{\theta}|x)}{p(\boldsymbol{\theta})}\right] p(\boldsymbol{\theta}|x) d\boldsymbol{\theta}\,.$$

(iii) Informative prior

An informative prior contains specific and definite information about the parameters in the model. This information can be determined from previous experiments or from an expert. This informative prior can also come from the literature or explicitly from an earlier posterior distribution.

2.4.3 Markov chain Monte Carlo (MCMC) methods

In this section, the definition of Markov chain will be introduced based on Robert and Casella (2004), Gelman et al. (1995), (Brooks et al., 2011) and Wakefield (2013). Moreover, we review the well-known ergodic theorem in this section. This theorem provides conditions for a Markov chain to work for Monte Carlo integration.

2.4.3.1 Markov chain

Let X be a random variable with state space $\boldsymbol{\chi}$, and let $\mathcal{B}(\boldsymbol{\chi})$ be the $\sigma-$algebra of $\boldsymbol{\chi}$. According to Robert and Casella (2004) and Gelman et al. (1995), the definitions for a Markov chain are presented as follows

Definition 1: A transition kernel is a function K defined on $\boldsymbol{\chi} \times \mathcal{B}(\boldsymbol{\chi})$ such that

i) $\forall x \in \boldsymbol{\chi}$, $K(x,.)$ is a probability measure

ii) $\forall A \in \mathcal{B}(\boldsymbol{\chi})$, $K(.,A)$ is measurable.

Definition 2: Given a transition kernel K, a sequence $\theta_0, \theta_1, ..., \theta_t, ...$ of random variables is a Markov chain if, for any t, the conditional distribution of θ_{t+1} given $\theta_0 = x_0, \theta_1 = x_1, ..., \theta_{t-1} = x_{t-1}, \theta_t = x_t$ equals the conditional distribution of θ_t given $\theta_t = x_t$. That is

$$\begin{aligned} p(\theta_{t+1} \in A|\theta_0 = x_0, \theta_1 = x_1, ..., \theta_{t-1} = x_{t-1}, \theta_t = x_t) &= p(\theta_t \in A|\theta_t = x_t) \\ &= \int_A K(x_t, x)dx,\ for\ \forall t. \end{aligned}$$

2.4.3.2 Ergodic theorem for Markov chains

If a Markov chain is ergodic, it converges to its unique stationary distribution. This allows the Markov chain to be sampled via the stationary distribution. The subsections below present the four conditions for a Markov chain to be ergodic (irreducibility, aperiodicity, positive recurrence, and reversibility) and the ergodic theorem itself.

Irreducibility:

Given a measure φ, the Markov chain (θ_t) with transition kernel $K(x, y)$ is $\varphi-$irreducible if, for every $A \in \mathcal{B}(\boldsymbol{\chi})$ with $\varphi(A) > 0$, there exists n such that $K^n(x, A) > 0$ for all $x \in \boldsymbol{\chi}$.

Aperiodicity:

The period of a state x is defined as

$$d(x) = g.c.d\{m \geq 1; K^m(x, x) > 0\}\,,$$

where g.c.d is the greatest common denominator. A Markov chain is aperiodic if all states have period 1.

Positive recurrence:

A Markov chain is recurrent if the chain will return to every state in a finite number of steps, with probability 1.

Reversibility:

A Markov chain with transition kernel K satisfies the detailed balance condition if there exists a probability density function f such that

$$K(y, x)f(y) = K(x, y)f(x)\,.$$

Then the density f is the invariant density of the chain and the chain is reversible.

Ergodic theorem: If (θ_t) is aperiodic, irreducible, positive recurrent with invariant distribution f then

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta_t) \to \int_{\theta} g(\theta)f(\theta)d\theta \quad \text{as } T \to \infty\,.$$

This theorem guarantees that a Markov chain satisfying the three conditions will converge to its unique stationary distribution. The following section introduces algorithms to sample a Markov chain from a stationary distribution.

2.4.3.3 MCMC algorithms

A MCMC method is used to sample for the parameters from a target distribution f. The target distribution is built from the conditional posterior distribution which has either a standard form or a non-standard form. The MCMC method produces an ergodic Markov chain (θ_t) with a stationary distribution f. After iterating until the chain converges, we can then use the chain to produce samples for the parameters.

In this section, we present the two well-known algorithms in MCMC, namely the Gibbs sampler (GS) algorithm and the Metropolis Hastings (MH) algorithm. The GS algorithm generates a Markov chain from the distribution of each parameter conditioned on the current estimated value of the other parameters. This means that the GS algorithm is used in the case that the conditional posterior distribution has a standard form. If the conditional posterior distribution does not have a standard form or is difficult to sample from, then we use the MH algorithm.

Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ be the vector of parameters in the model and let $\boldsymbol{y}$ be the observed data. According to Geman and Geman (1984), Gelman et al. (1995) and Brooks et al. (2011), the GS algorithm produces a Markov chain associated with the standard posterior distribution f through the following steps:

Step 1: Initialise $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$

Step 2: For $t = 1$ to T do

2.1: Sample $\theta_1^{(t)}$ from $f(\theta_1|\theta_2^{(t-1)}, ..., \theta_p^{(t-1)}, \boldsymbol{y})$

2.2: Sample $\theta_2^{(t)}$ from $f(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}..., \theta_p^{(t-1)}, \boldsymbol{y})$

.....

2.3: Sample $\theta_p^{(t)}$ from $f(\theta_p|\theta_1^{(t)}, \theta_2^{(t)}..., \theta_{p-1}^{(t)}, \boldsymbol{y})$

Step 3: End for

Another MCMC algorithm is the MH algorithm introduced by Metropolis et al. (1953) and Hastings (1970). This algorithm produces a Markov chain associated with the target distribution f and the proposal distribution q. Here, the target distribution f does not

have a standard form. This algorithm becomes more complex when the dimension of the parameter vector is large. The MH algorithm for a parameter of one dimension is presented through the following steps:

Step 1: Initialise $\theta^{(0)}$

Step 2: For $t = 1$ to T do

Step 3: Given the current parameter value $\theta = \theta^{(t)}$, propose θ' from the proposal dsitribution $q(\theta'|\theta^{(t)})$

Step 4: Calculate acceptance probability $\alpha(\theta^{(t)} \rightarrow \theta') = \frac{f(\theta')q(\theta^{(t)}|\theta')}{f(\theta^{(t)})q(\theta'|\theta^{(t)})}$

Step 5: Simulate u from uniform distribution from 0 to 1, $u \sim \mathcal{U}(0,1)$

5.1: If $\alpha(\theta^{(t)} \rightarrow \theta') > u$, then set $\theta^{(t+1)} = \theta'$

5.2: Else, set $\theta^{(t+1)} = \theta^{(t)}$

Step 6: End for

From the two algorithms, we can see that the GS algorithm is a simple case of MH when the acceptance rate is equal 1. Finally, to prove that the Markov chains produced by the MH algorithm or the GS algorithm are reversible and converge to the target distribution, Robert and Casella (2004) presented the following theorem.

Theorem: Let (θ_t) be the chain produced by the Metropolis-Hastings algorithm. For every conditional posterior distribution f

a) The kernel of the chain satisfies the detailed balance condition with f.

b) f is a stationary distribution of the chain.

2.4.3.4 Choices for the proposal distribution.

There are several ways to choose the proposal distribution to ensure that the whole parameter space is explored. In this section, we will review the two popular choices of the proposal distribution for a continuous parameter state space (Robert and Casella, 2004;

Wakefield, 2013; Gelman et al., 1995). They are a symmetric random walk Metropolis-Hastings and an independent Metropolis-Hastings algorithm.

Random walk Metropolis-Hastings

A random walk is a sequence of random variables (θ_t) which satisfies

$$\theta_{t+1} = \theta_t + \epsilon_t \,,$$

where ϵ_t is generated independently of $\theta_{t+1}, \theta_t, ..., \theta_0$. If the distribution of ϵ_t is symmetric about 0, the sequence is called a symmetric random walk.

The symmetric random walk is the most popular choice of a proposal distribution for a continuous parameter, especially a normal random walk. This is because this proposal distribution satisfies the following properties (Gelman et al., 1995):

i) It is easy to propose a new value from a current value;

ii) It is easy to compute the acceptance rate. In particular, the acceptance rate for a normal random walk has the form

$$\alpha(\theta_t \to \theta_{t+1}) = \min\left\{\frac{f(\theta_{t+1})}{f(\theta_t)}, 1\right\} ;$$

iii) The distance for each jump depends on the variance of a proposal distribution.

Independent Metropolis-Hastings

The proposed parameter value does not depend on the current state of the chain (Gelman et al., 1995). If we denote the proposed distribution q following the distribution $q \sim g(y)$, then the acceptance probability in the MH algorithm is calculated as

$$\alpha(\theta_t \to \theta_{t+1}) = \min\left\{\frac{f(\theta_{t+1})g(\theta_t)}{f(\theta_{t+1})g(\theta_{t+1})}, 1\right\} .$$

The convergence properties of the MH chain follow the properties of the distribution g.

Chapter 3

Penalized Spline Joint Models for Longitudinal and Time-to-event Data: An ECM Approach

3.1 Introduction

Joint models for longitudinal data and time-to-event data are aimed to measure the association between the longitudinal marker level and the hazard rate for an event. Longitudinal data are collected repeatedly for several subjects. In this data, there are two types of covariates, namely, time-independent covariates and time-dependent covariates. Furthermore, there are two different categories of time-dependent covariates, namely, external and internal covariates. In clinical studies, internal time-dependent longitudinal outcomes are often applied to monitor disease progression and failure time.

In modern survival analysis, the Cox models have been a very popular joint model for time-independent covariates (Cox, 1972). These models measure the effect of time-independent covariates on the hazard rate for an event. Subsequently, the extended Cox models were developed for external time-dependent covariates. However, these latter models cannot handle longitudinal biomarkers. Therefore, Rizopoulos (2012) introduced joint models for internal time-dependent covariates and the risk for an event based on linear mixed-effects models and relative risk models.

The basic assumption for the standard joint models proposed by Rizopoulos (2012) is that the hazard rate at a given time of the dropout process is associated with the expected value of the longitudinal responses at the same time. The whole history of longitudinal responses has an influence on the survival function. Thus, it is crucial to obtain good

estimates for the subject-specific trajectories in order to have an accurate estimation of the survival function. In addition, an important feature we need to account for is that many observations in the sample often show non-linear and fluctuated longitudinal trajectories in time. Each observation has its own trajectory. Therefore, flexibility is needed for subject-specific longitudinal submodels in the joint models to improve the predictions.

There are several previous efforts to model flexibly the subject-specific longitudinal profiles in the joint models. Brown et al. (2005) applied B-splines with multidimensional random effects. In particular, Brown et al. (2005) assumed that both subject and population trajectories have the same number of basis functions. By doing this, the number of parameters in the longitudinal submodel is reasonably large. If we have to deal with the roughness of the fit for this model, the computational problems increase especially when the dimension of the random effects vector is large. Ding and Wang (2008) proposed the use of B-splines with a single multiplicative random effect, to link the population mean function with the subject-specific profile. This simple model allows an easy estimation of parameters, however it may not be appropriate for many practical applications (Rizopoulos, 2011). Rizopoulos (2011); Crowther and Lambert (2013) have considered more flexible models using natural cubic splines with the expansion of the random effects vector. The roughness of the fit is not mentioned in these models.

The original contributions in this chapter include the new approaches to model non-linear shapes of subjects-specific evolutions for joint models by extending the standard joint models of Rizopoulos (2012). In particular, we implement penalized splines using a truncated polynomial basis for the longitudinal submodel. Following this, the linear mixed effects approach is applied to model the individual trajectories and impose smoothness over adjacent coefficients respectively. The ECM algorithm for this proposed model is designed for parameter estimation. In addition to this, corresponding standard errors are calculated using the observed information matrix. However, as the matrices of random effects covariates in our models are different from the matrices of random effects covariates in the standard joint models, the JM package of Rizopoulos (Rizopoulos, 2010) cannot be used for our models. Therefore, a set of R codes – as part of the original contributions – was written for the penalized spline joint models to implement the proposed procedures on the extensive simulation studies and case studies.

This chapter is organized as follows. Section 3.2 describes the penalized splines with trun-

cated polynomial basis for the joint models. In this section, the two models are specified as a penalized spline joint model with hazard rate at baseline having Gompertz distribution (referred to as Model 1) and a penalized spline joint model with a piecewise-constant baseline risk function (referred to as Model 2). The joint likelihood, score functions and the ECM algorithm for estimation are presented in Section 3.3. Then we validate the proposed algorithm using extensive simulation studies and apply it to a case study based on AIDS data in Section 3.4. Finally, Section 3.5 gives concluding remarks.

3.2 The penalized spline joint models

In this section, we introduce the joint models using penalized splines with truncated polynomial basis. The proposed parametrization is based on the standard joint models of Rizopoulos (2012) and the regression model of a longitudinal response using penalized splines.

The notation in this section is taken from Rizopoulos (2012). Let T_i^* be the true survival time and C_i be the censoring time for the i^{th} subject ($i = 1, ..., n$). T_i denotes the observed failure time for the i^{th} subject ($i = 1, ..., n$), which is defined as $T_i = \min(T_i^*, C_i)$. If an i^{th} subject is not censored, this means that we have observed its survival time, so $T_i \leq C_i$. If an i^{th} subject is censored, this means that we lose its follow up, or the subject has died from other causes, in this case $T_i > C_i$. Furthermore, we define the event indicator as $\delta_i = I(T_i^* \leq C_i)$. In this thesis, we only consider noniformative right censoring. The observed data for survival outcome are (T_i, δ_i), $i = 1, ..., n$.

For a longitudinal response, suppose that we have n subjects in the sample and the actual observed longitudinal data for the i^{th} subject at time point t is $y_i(t)$. We measure the i^{th} subject at n_i time points. Thus, the longitudinal data consists of the measurements $y_{ij} = \{y_i(t_{ij}),\ j = 1, ..., n_i\}$ taken at time points t_{ij}. We denote the true and unobserved value of the longitudinal outcome at time t as $m_i(t)$. We assume that the relation between $y_i(t)$ and $m_i(t)$ is

$$y_i(t) = m_i(t) + \varepsilon_i(t)\,, \tag{3.2.1}$$

where $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

In penalized spline regression models (Ruppert et al., 2003; Durban et al., 2005), the observed longitudinal covariate is modelled using truncated power functions with a general power basis of degree p. Moreover, the longitudinal response is also parameterized as a linear mixed effects model to specify the individual curves and impose the amount of smoothing. As a result, the coefficients of the knots will be constrained to handle smoothing. In particular, the longitudinal submodel for the i^{th} subject at time point t_{ij} is

$$\begin{aligned} y_{ij} &= f(t_{ij}) + g_i(t_{ij}) + \varepsilon_i(t_{ij}),\ \varepsilon_i(t_{ij}) \sim \mathcal{N}(0, \sigma_\varepsilon^2), \\ f(t_{ij}) &= \beta_0 + \beta_1 t_{ij} + ... + \beta_p t_{ij}^p + \sum_{k=1}^{K} u_{pk}(t_{ij} - \mathcal{K}_k)_+^p, \\ g_i(t_{ij}) &= v_{i0} + v_{i1} t_{ij} + v_{i2} t_{ij}^2 + ... + v_{ip} t_{ij}^p + \sum_{k=1}^{K} w_{ipk}(t_{ij} - \mathcal{K}_k)_+^p \,. \end{aligned} \tag{3.2.2}$$

Here, the set $\left\{1, t_{ij}, ..., t_{ij}^p, (t_{ij} - \mathcal{K}_1)_+^p, ..., (t_{ij} - \mathcal{K}_K)_+^p\right\}$ is known as the truncated power basis of degree p. Moreover, $\mathcal{K}_1, ..., \mathcal{K}_K$ are K fitted knots, chosen following Ruppert et al. (2003), Chapter 5. The function $f(.)$ is the smooth function which reflects the overall trend of the population. The functions $g_i(.)$ are the smooth functions that reflect the individual curves. To constrain the coefficient of knots, the vector $(u_{p1}, ..., u_{pK})^T$ in the function $f(.)$ is treated as random effects. Therefore, $\boldsymbol{\beta}^T = (\beta_0, ..., \beta_p)$ is a $((p+1)\times 1)$ row vector of fixed effects and $\boldsymbol{b}_i^T = (u_{p1}, ..., u_{pK}, v_{i0}, ..., v_{ip}, w_{ip1}, ..., w_{ipK})$ is a $((p+2K+1)\times 1)$ vector of random effects for the i^{th} subject. The assumptions for the random effects for the i^{th} subject are $(v_{i0}, ..., v_{ip})^T \sim \mathcal{MVN}(0, \sum)$, $u_{pk} \sim \mathcal{UVN}(0, \sigma_u^2)$, $w_{ipk} \sim \mathcal{UVN}(0, \sigma_w^2)$ and they are independent of one another. We can now rewrite (3.2.2) as

$$\begin{aligned} y_i(t_{ij}) &= f(t_{ij}) + g_i(t_{ij}) + \varepsilon_i(t_{ij}) \\ &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + ... + \beta_p t_{ij}^p + \sum_{k=1}^{K} (u_{pk} + w_{ipk})(t_{ij} - \mathcal{K}_k)_+^p \\ &\quad + v_{i0} + v_{i1} t_{ij} + v_{i2} t_{ij}^2 + ... + v_{ip} t_{ij}^p + \varepsilon_i(t_{ij}) \,. \end{aligned} \tag{3.2.3}$$

Let $u_{ipk} = u_{pk} + w_{ipk}$ and note that $u_{ipk} \sim \mathcal{UVN}(0, \sigma_u^2 + \sigma_w^2)$. In order to allow greater flexibility, we assume that $(u_{ip1}, ..., u_{ipK})^T \sim \mathcal{MVN}(0, \boldsymbol{D})$, where $\boldsymbol{D} = Diag(D_{11}, ..., D_{KK})$. By doing this, the dimension of the vector of random effects, $\boldsymbol{b}_i^T = [\ v_{i0}, ..., v_{ip} \quad u_{ip1}, ..., u_{ipK}\ ]$, decreases to $((p + K + 1) \times 1)$. Consequently, the dimension of the multi-integrals in the log-likehood function in (3.3.2) also decreases. This presentation is crucial for reducing

the computational problems while coding. The model in (3.2.3) now becomes:

$$\begin{aligned} y_i(t_{ij}) &= f(t_{ij}) + g_i(t_{ij}) + \varepsilon_i(t_{ij}) \\ &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + ... + \beta_p t_{ij}^p + \sum_{k=1}^{K} u_{ipk}(t_{ij} - \mathcal{K}_k)_+^p \\ &\quad + v_{i0} + v_{i1} t_{ij} + v_{i2} t_{ij}^2 + ... + v_{ip} t_{ij}^p + \varepsilon_i(t_{ij})\,. \end{aligned} \tag{3.2.4}$$

The model in (3.2.4) can be rewritten in matrix notation as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}\,, \tag{3.2.5}$$

where

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix}, \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X}_1 & 0 & \dots & 0 & \boldsymbol{Z}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{X}_2 & \dots & 0 & 0 & \boldsymbol{Z}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{X}_n & 0 & 0 & \dots & \boldsymbol{Z}_n \end{bmatrix},$$

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & \cdots & t_{i1}^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & \cdots & t_{in_i}^p \end{bmatrix}, \boldsymbol{Z}_i = \begin{bmatrix} (t_{i1} - \mathcal{K}_1)_+^p & \cdots & (t_{i1} - \mathcal{K}_K)_+^p \\ \vdots & \vdots & \vdots \\ (t_{in_i} - \mathcal{K}_1)_+^p & \cdots & (t_{in_i} - \mathcal{K}_K)_+^p \end{bmatrix},$$

$$\boldsymbol{b}^T = (v_{10}, ..., v_{1p}, ..., v_{n0}, ..., v_{np}, u_{1p1}, ..., u_{1pK}, ..., u_{np1}, ..., u_{npK})\,,$$

$$\boldsymbol{\beta}^T = (\beta_0, ..., \beta_p)\,.$$

Here, $\boldsymbol{y}$ is the $\left(\sum_{i=1}^{n} n_i \times 1\right)$ matrix of observed longitudinal data; $\boldsymbol{X}$ is the $\left(\sum_{i=1}^{n} n_i \times (p+1)\right)$ matrix of fixed effect covariates; $\boldsymbol{Z}$ is the $\left(\sum_{i=1}^{n} n_i \times (p+K+1)n\right)$ matrix of random effect covariates and $\boldsymbol{\varepsilon}$ is the $\left(\sum_{i=1}^{n} n_i \times 1\right)$ matrix of error.

Postulating a proportional hazard model, the penalized spline joint model for longitudinal and time to event data is defined by

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= \lim_{dt \to 0} Pr\left\{t \le T_i^* < t + dt \mid T_i^* \ge t, \mathcal{M}_i(t), \boldsymbol{w}_i\right\}/dt \\ &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}\,, \end{aligned} \tag{3.2.6}$$

where $h_0(t)$ is the hazard at baseline and $\boldsymbol{w}_i$ is a vector of baseline covariates (such as treatment indicator, gender of a patient, etc). Furthermore, $\mathcal{M}_i(t) = \{m_i(s),\, 0 \le s < t\}$ denotes the history of the true unobserved longitudinal process up to time point t.

Using (3.2.5), the longitudinal submodel for the i^{th} subject is given by

$$\begin{cases} y_i(t) & = m_i(t) + \varepsilon_i(t),\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ m_i(t) & = \boldsymbol{X}_i^T(t)\boldsymbol{\beta} + \boldsymbol{X}_i^T(t)\boldsymbol{v}_i + \boldsymbol{Z}_i^T(t)\boldsymbol{u}_i \\ & \boldsymbol{v}_i \sim \mathcal{MVN}(\mathbf{0}, \sum),\ \boldsymbol{u}_i \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{D})\,, \end{cases} \tag{3.2.7}$$

where the covariance matrix of random effects $\boldsymbol{b}_i^T = (v_{i0}, ..., v_{ip}, u_{ip1}, ..., u_{ipK})$ is given as

$$\boldsymbol{G} = Cov(\boldsymbol{b}_i) = \begin{bmatrix} \sum & 0 \\ 0 & \boldsymbol{D} \end{bmatrix}.$$

To complete the specification of the model in (3.2.6), we now need to define the form for the baseline risk function $h_0(.)$. Motivated by the fact that in real life, $h_0(.)$ is usually unknown. Two options are adopted to determine the form of the function $h_0(.)$ in this chapter. Firstly, a standard option is to use a known parametric distribution for the risk function. For this option, the Gompertz distribution is chosen. Secondly, the piecewise constant model is chosen when $h_0(.)$ is considered completely unspecified.

Therefore, the proposed penalized spline joint models considered in this chapter are as follows:

Model 1: Penalized spline joint model with hazard rate at baseline having a Gompertz distribution

$$\begin{cases} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) = \lambda_1 \exp(\lambda_2 t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\} \\ m_i(t) = \boldsymbol{X}_i^T(t)\boldsymbol{\beta} + \boldsymbol{X}_i^T(t)\boldsymbol{v}_i + \boldsymbol{Z}_i^T(t)\boldsymbol{u}_i\,. \end{cases} \tag{3.2.8}$$

Model 2: Penalized spline joint model with a piecewise-constant baseline risk function

$$\begin{cases} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) = \sum_{q=1}^{Q} \xi_q I(\nu_{q-1} < t \leq \nu_q) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\} \\ m_i(t) = \boldsymbol{X}_i^T(t)\boldsymbol{\beta} + \boldsymbol{X}_i^T(t)\boldsymbol{v}_i + \boldsymbol{Z}_i^T(t)\boldsymbol{u}_i\,, \end{cases} \tag{3.2.9}$$

where $0 = \nu_0 < \nu_1 < ... < \nu_Q$ denotes a split of the time scale, with ν_Q being larger than the largest observed time, and ξ_q denotes the value of the baseline hazard in the interval $[\nu_{q-1}, \nu_q)$. In both models, $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, $\boldsymbol{\beta}$, $\boldsymbol{v}_i$ and $\boldsymbol{u}_i$ are given as in (3.2.5).

3.3 Parameter estimation

After defining the two penalized spline joint models in Section 3.2, we present the joint likelihood and score functions for the parameters in the models. Following this, the ECM algorithm is explained in detail.

3.3.1 Likelihood and score functions

Following Rizopoulos (2012), both the longitudinal and survival processes contain the vector of time-independent random effects b_i. This means that

$$\begin{aligned} p(T_i, \delta_i, \boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) &= p(T_i, \delta_i|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) \\ p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) &= \prod_j p\left\{\boldsymbol{y_i}(t_{ij})|\boldsymbol{b_i}; \boldsymbol{\theta}\right\}, \end{aligned} \tag{3.3.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_t^T, \boldsymbol{\theta}_y^T, \boldsymbol{\theta}_b^T)^T$ denotes the full parameter vector with $\boldsymbol{\theta}_t = (\gamma^T, \alpha, \theta_{h_0}^T)^T$ denoting the parameter vector for the survival outcomes. Furthermore, $\boldsymbol{\theta}_y = (\boldsymbol{\beta}^T, \sigma_\varepsilon^2)^T$ is the parameter vector for longitudinal outcomes and $\boldsymbol{\theta}_b = vech(\boldsymbol{G})$ is the vector-half of the variance matrix of random effects. In addition, we assume that the hazard rate at time t, conditional on the covariate path, depends on the current value of longitudinal outcomes. The censoring mechanism is independent of the true event times and future longitudinal measurements. Under these assumptions, the log-likelihood formulation of the penalized spline joint models can be written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= l(\boldsymbol{\theta} \mid T_i, \delta_i, \boldsymbol{y}_i) \\ &= \sum_i \log \int_{\boldsymbol{b}_i} p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i, \end{aligned} \tag{3.3.2}$$

where the conditional density for the survival part has the form

$$\begin{aligned} p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) &= h(T_i \mid \mathcal{M}_i(T_i), \boldsymbol{w}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta})^{\delta_i} S(T_i \mid \mathcal{M}_i(T_i), \boldsymbol{w}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) \\ &= \left[h_0(T_i) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(T_i)\right\}\right]^{\delta_i} \exp\left[-\int_0^{T_i} h_0(s) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s) ds\right\}\right]. \end{aligned} \tag{3.3.3}$$

Moreover, the density for the longitudinal part with the random effects is given by

$$\begin{aligned} p(\boldsymbol{y_i}|\boldsymbol{b}_i;\boldsymbol{\theta}_y)p(\boldsymbol{b_i};\boldsymbol{\theta}_b) &= \prod_j p\left\{y_i(t_{ij})|\boldsymbol{b_i};\boldsymbol{\theta}_y\right\} p(\boldsymbol{b_i};\boldsymbol{\theta}_b) \\ &= \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{n_i}{2}}} \exp\left\{-\frac{\parallel y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i \parallel^2}{2\sigma_\varepsilon^2}\right\} \\ &\quad \times (2\pi)^{-\frac{q_b}{2}} \det(\boldsymbol{G})^{-1/2} \exp(-\boldsymbol{b}_i^T\boldsymbol{G}^{-1}\boldsymbol{b_i}/2)\,, \end{aligned} \tag{3.3.4}$$

where q_b denotes the dimensionality of the random effects vector.

We consider the log-likelihood of (T_i, δ_i, $\boldsymbol{y}_i$, $\boldsymbol{b}_i$) over the unknowns $\boldsymbol{\theta}_t$, $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_b$

$$\log l(\boldsymbol{\theta} \mid T_i, \delta_i, \boldsymbol{y}_i, \boldsymbol{b_i}) = \log p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) + \log p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\beta}) + \log p(\boldsymbol{b}_i; \boldsymbol{G})\,.$$

The function for maximizing the log-likelihood involves the density function of the survival time and the least squares with a penalty term, which is

$$\log p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) - \frac{(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{X}_i\boldsymbol{v}_i - \boldsymbol{Z}_i\boldsymbol{u}_i)^T(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{X}_i\boldsymbol{v}_i - \boldsymbol{Z}_i\boldsymbol{u}_i)}{\sigma_\varepsilon^2} - \boldsymbol{b}_i^T\boldsymbol{G}^{-1}\boldsymbol{b_i}\,. \tag{3.3.5}$$

According to Ruppert et al. (2003), the term $\sigma_\varepsilon^2\boldsymbol{b}_i^T\boldsymbol{G}^{-1}\boldsymbol{b_i}$ is called a roughness penalty and the variance components matrix is defined as $\boldsymbol{F} = \sigma_\varepsilon^2\boldsymbol{G}^{-1}$. Using a Lagrange multiplier argument, the variance components matrix is the condition to constrain the coefficients of the knots $\boldsymbol{u}_i$. These restrict the influence of the variables $(t - K_k)_+^p$ and lead to smoother spline functions.

Using (3.3.2), the score vector for the penalized spline joint models can be expressed as:

$$\begin{aligned} S(\boldsymbol{\theta}) &= \sum_i \frac{\partial}{\partial\boldsymbol{\theta}^T} \log \int p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i \\ &= \sum_i \int \frac{\partial}{\partial\boldsymbol{\theta}^T} \log \left\{p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\right\} . p(\boldsymbol{b}_i \mid T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i\,. \end{aligned} \tag{3.3.6}$$

The requirement for numerical integration with respect to the random effects is one of the main difficulties in the joint models (Rizopoulos, 2012). The maximum likelihood estimation (MLEs) is typically obtained using standard maximisation algorithms such as the expectation maximization algorithm or the Newton-Raphson algorithm.

3.3.2 The ECM algorithm

The EM algorithm has been widely used in joint models, such as the standard joint model of Rizopoulos (Rizopoulos, 2012) and the generalised linear mixed joint model (Viviani et al., 2014). The ECM algorithm is a natural extension of the EM algorithm for which the maximisation process on the M-step is conditional on some functions of the parameters under estimation. It can also reduce computer time. The ECM algorithm following McLachlan and Krishnan (2007) will be used to obtain the maximum likelihood estimates of the penalized spline joint models in this chapter.

In these models, the random effects $\boldsymbol{b}_i$ are considered to be missing data. Hence, it is difficult to estimate directly the parameter vector $\boldsymbol{\theta}$ that maximizes the observed data log-likelihood $l(\boldsymbol{\theta})$ in (3.3.2). Alternatively, we can estimate the parameter vector $\boldsymbol{\theta}$ that maximizes the expected value of the complete data log-likelihood which is

$$E\left\{\log p(T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b}_i; \boldsymbol{\theta}) | T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}^{(it)}\right\},$$

where $\boldsymbol{\theta}^{(it)}$ is the parameter vector given at the i^{th} iteration.

The following are the steps of this algorithm.

Step 1: *Initialization*

First initialise the parameters. We assume that there are m parameters in the models and the starting value of the parameter vector is $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_m^{(0)})$. Based on these initial values, we calculate the log-likelihood using (3.3.2).

Step 2: *The E-step for the penalized joint models*

Fill in the missing data and replace the log-likelihood function of the observed data with the expected function of the complete data log-likelihood as follows

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(it)}) &= \sum_i \int \log\left\{p(T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b}_i; \boldsymbol{\theta})\right\} . p(\boldsymbol{b}_i | T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i \\ &= \sum_i \int \left(\log p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{y_i} | \boldsymbol{b}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{b}_i; \boldsymbol{\theta})\right) . p(\boldsymbol{b}_i | T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i \,. \end{aligned} \tag{3.3.7}$$

Step 3: *The conditional M-step for the penalized joint models*

This step is implemented in 4 stages as follows:

3.1 Given the current value of the parameter vector at the i^{th} iteration $\boldsymbol{\theta}^{(it)} = (\theta_1^{(it)}, \theta_2^{(it)}, ..., \theta_m^{(it)})$, calculate the log-likelihood $l(\boldsymbol{\theta}^{(it)}) = \sum_i \log \int_{b_i} p(T_i, \delta_i, \boldsymbol{y}_i, \boldsymbol{b}_i; \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i$.

3.2 Propose the new value for the first parameter $\theta_1^{(prop)}$ which maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(it)})$. Then calculate the log-likelihood $l(\boldsymbol{\theta}^{(prop)})$ where $\boldsymbol{\theta}^{(prop)} = (\theta_1^{(prop)}, \theta_2^{(it)}, ..., \theta_m^{(it)})$.

3.3 Set $\boldsymbol{\theta}_1^{(it)} = \boldsymbol{\theta}^{(prop)}$ if $l(\boldsymbol{\theta}^{(prop)}) \geq l(\boldsymbol{\theta}^{(it)})$, otherwise set $\boldsymbol{\theta}_1^{(it)} = \boldsymbol{\theta}^{(it)}$.

3.4 Similarly, based on the value of the parameter vector $\boldsymbol{\theta}_1^{(it)}$, update the new value for the second parameter, continue updating until the last parameter, $\boldsymbol{\theta}_m^{(it)}$, and then set $\boldsymbol{\theta}^{(it+1)} = \boldsymbol{\theta}_m^{(it)}$.

Step 4: *Iterate steps 2-3 until the algorithm converges numerically.*

To update the new values for parameters in the conditional M-step, we have the closed-form estimates for the measurement error variance σ_ε^2 and the covariance matrix of the random effects respectively by maximizing the expected function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(it)})$. Unfortunately, we cannot obtain closed-form expressions for the remaining parameters. Thus we employ the one-step Newton-Raphson approach to get the updates for $\boldsymbol{\beta}^{(it+1)}$, $\boldsymbol{\gamma}^{(it+1)}$, $\alpha^{(it+1)}$ and $\boldsymbol{\theta}_{h_0}^{(it+1)}$ respectively as detailed in Appendix A.2.

Following Louis (1982), standard errors for the parameter estimates can be calculated by using the estimated observed information matrix

$$\hat{var}(\hat{\boldsymbol{\theta}}) = \left\{\mathcal{I}(\hat{\boldsymbol{\theta}})\right\}^{-1},$$

where

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n} \frac{\partial S_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

3.4 Empirical results

This section presents two simulation studies for Model 1 in (3.2.8), whereas Model 2 in (3.2.9) will be applied to a case study only. In Section 3.4.1, we simulate data from Model 1 with three internal knots in the longitudinal submodel and Gompertz distribution

for the baseline risk function. In Section 3.4.2, we simulate data from Model 1 having Gompertz distribution for the baseline risk function and non-linear logarithm subject-specific trajectories. The ECM algorithm, written in R code, is applied to estimate the true values of parameters in both cases. These procedures for Model 1 and Model 2 are then applied to AIDS data in Section 3.4.3. Model comparison between Model 1, Model 2 and Rizopoulos's standard joint model (as Model 3) for the AIDS data will be presented at the end of this section.

3.4.1 Simulation study 1

3.4.1.1 Data description

Recall the penalized spline joint Model 1 of (3.2.8) with three internal knots in longitudinal submodel and the Gompertz distribution for the baseline risk function in the form

$$h_i(t) = h_0(t)exp(\gamma x_i + \alpha(m_i(t))) = \lambda_1 exp(\lambda_2 t)exp\left\{\gamma x_i + \alpha m_i(t)\right\} . \tag{3.4.1}$$

Here $h_0(t)$ is the hazard function at baseline having Gompertz distribution, x_i is the baseline covariate and $m_i(t)$ denotes the true and unobserved value of the longitudinal at time t. The form of $m_i(t)$ is given by

$$m_i(t) = \beta_0 + \beta_1 t + u_{i1}(t - \mathcal{K}_1)_+ + u_{i2}(t - \mathcal{K}_2)_+ + u_{i3}(t - \mathcal{K}_3)_+ + v_{i0} , \tag{3.4.2}$$

where $\boldsymbol{b}_i = (u_{11}, u_{12}, u_{13}, v_{i0})^T$ is the vector of random effects and is assumed to have a normal distribution with mean zero and diagonal covariance matrix $\boldsymbol{D} = Diag(D_{11}, D_{22}, D_{33}, D_{44})$. $\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3$ denote the three internal knots put into the model. The observed longitudinal value at time point t for the i^{th} subject is of the form

$$y_i(t) = m_i(t) + \varepsilon_i(t) , \tag{3.4.3}$$

where the error variable $\varepsilon_i(t)$ is assumed to come from a normal distribution with mean zero and variance σ^2.

The vector of all the parameters $\boldsymbol{\theta}$ for the models in (3.4.1) and (3.4.2) is $\boldsymbol{\theta} = (\boldsymbol{\theta}_t^T, \boldsymbol{\theta}_y^T, \boldsymbol{\theta}_b^T)^T$, where $\boldsymbol{\theta}_t = (\gamma, \alpha, \lambda_1, \lambda_2)^T$ denotes the parameter vector for the survival outcomes. Furthermore, $\boldsymbol{\theta}_y = (\beta_0, \beta_1, \sigma_\varepsilon^2)^T$ is the parameter vector for longitudinal outcomes and $\boldsymbol{\theta}_b = \boldsymbol{D}$ is the variance matrix of random effects.

To simulate the observed survival time T_i of the joint model in (3.4.1), we applied the methods adapted by Bender et al. (2005), Austin (2012) and Crowther and Lambert (2013) to generate the true survival time. The detail of simulating survival time is presented in Appendix A.3. We further assumed that the censoring mechanism is exponentially distributed with parameter λ. The observed survival time was the minimum of the censoring time and the true survival time. We generated the survival time T_i for $n = 500$ subjects with the parameters: $\beta_0 = 5$, $\beta_1 = 2$, $\lambda_1 = 0.1$, $\lambda_2 = 0.5$, $\gamma = 0.5$,$\alpha = 0.05$, $\delta = 2$ and $D = Diag(2, 2, 2, 4)$. Then we generated the longitudinal responses $m_i(t)$ using (3.4.2). The simulated model is therefore

$$\begin{cases} h_i(t) = 0.1exp(0.5t)exp\left\{0.5x_i + 0.05m_i(t)\right\} \\ m_i(t) = 5 + 2t + u_{i1}(t-1)_+ + u_{i2}(t-2)_+ + u_{i3}(t-3)_+ + v_{i0}\,. \end{cases} \tag{3.4.4}$$

The sample of simulated data is presented in Appendix A.1. The curve of the Kaplan-Meier estimate for the survival function of simulated data (left panel) and the longitudinal trajectories for the whole simulated sample (right panel) are presented in Figure 3.1. The dashed lines in the left panel correspond to the 95% pointwise confidence intervals (CIs). It is clear from the plot of the Kaplan-Meier estimator that the survival probability starts from 1 and decreases gradually until the 5^{th} month of the study. It is nearly zero after 6 months or so. The right panel is the longitudinal trajectories for the first 100 subjects reflecting the form as in (3.4.2).

3.4.1.2 Parameter estimation

The ECM algorithm, as described in Section 3.3.2, was implemented to estimate all parameters in (3.4.1). The initial values of the parameters were set at $\beta_0 = 1$, $\beta_1 = 1$, $\lambda_1 = 0.05$, $\lambda_2 = 0.1$, $\gamma = 0.1$, $\alpha = 0.01$, $\sigma = 1$, $D_{11} = 3$, $D_{22} = 3$, $D_{33} = 3$, $D_{44} = 3$ respectively. However, these initial values can also be set randomly. The traces of each of these parameters are presented in Figures 3.2 and 3.3 respectively, which show how the algorithm updates new values of the parameters. They also demonstrate the convergence of the algorithm after 10 to 20 iterations. In particular, the parameters β_0, β_1, λ_2, α, σ, D_{22} and D_{33} converge linearly to the true values while the parameters λ_1, γ, D_{11}, and D_{44} oscillate before converging to the true values.

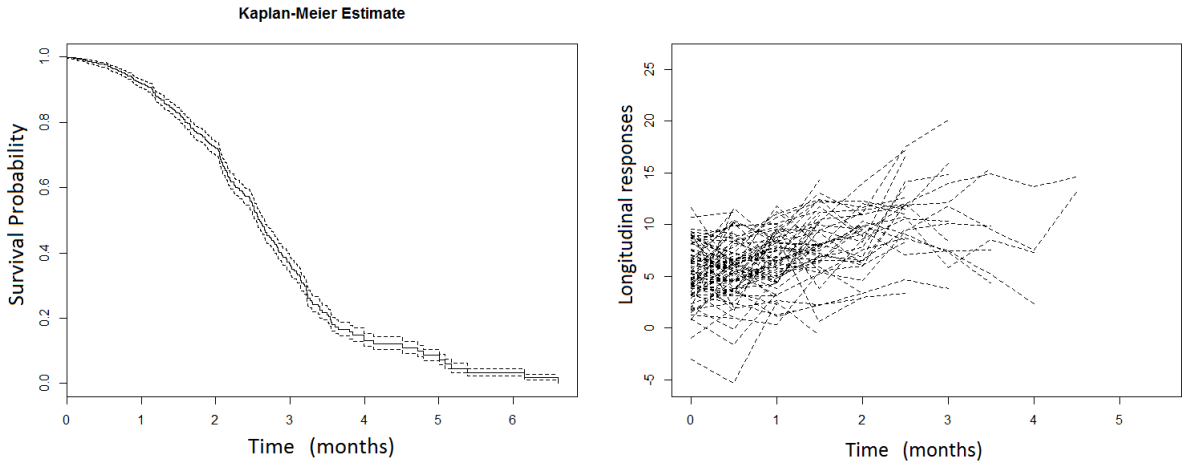

Figure 3.1: The Kaplan-Meier estimate of the survival function of the simulated data of (3.4.1) (left panel). Longitudinal trajectories of the first 100 subjects from the simulated sample of (3.4.2) (right panel).

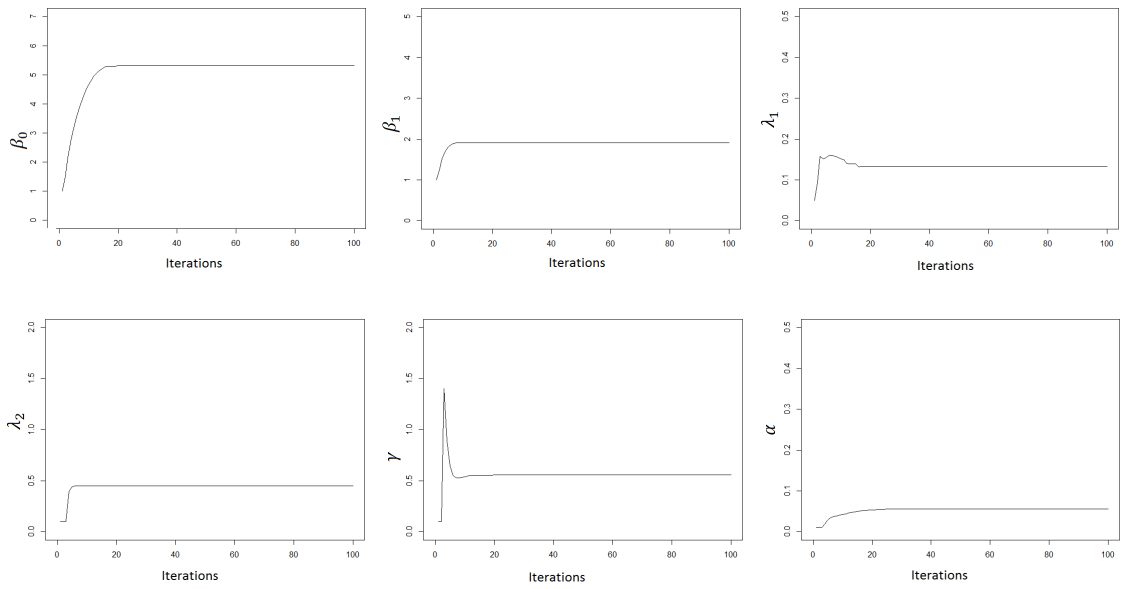

Figure 3.2: The traces plot of the parameters β_0, , β_1, λ, γ and α for 100 iterations.

We ran the simulation for 30 independent samples with different sample sizes (n = 200, 300, and 500). Then, the means, standard deviations (SD) and mean square errors (MSE) of parameters were calculated and are presented in Table 3.1. The point estimates of each parameter are reasonably close to the true values when the sample sizes are 300 and 500. This is also supported by the values of the SD and the MSE which

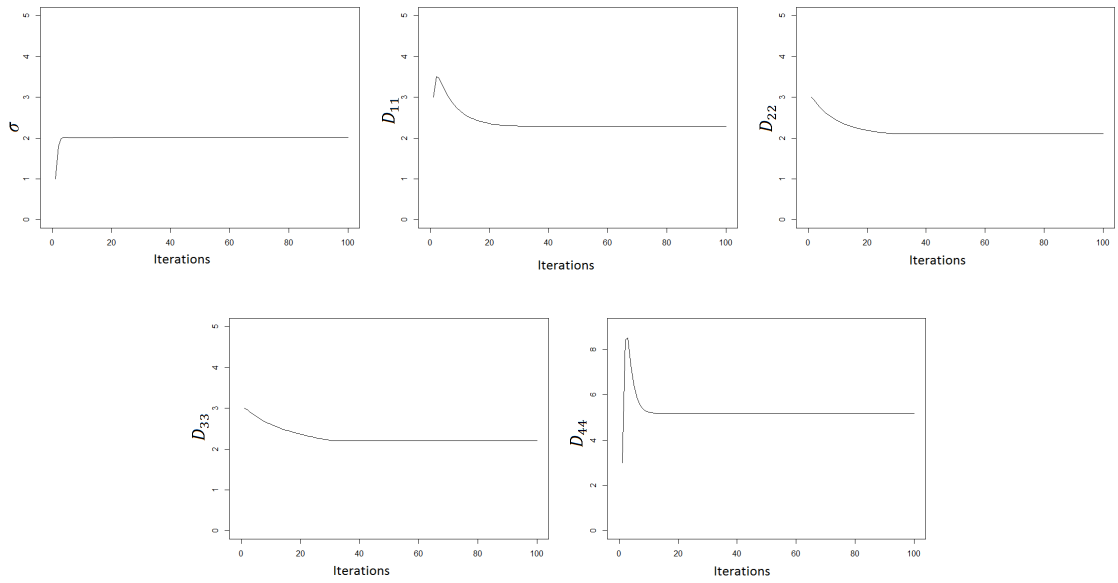

Figure 3.3: The traces of the parameters σ, D_{11}, D_{22}, D_{33}, D_{44} for 100 iterations.

Table 3.1: Summary statistics for parameter estimation of the simulated data of the model in (3.4.4) for different sample sizes.

Parameter	True value	$n = 200$			$n = 300$			$n = 500$		
		Estimate	SD	MSE	Estimate	SD	MSE	Estimate	SD	MSE
β_0	5	4.21	0.72	0.76	4.68	0.50	0.32	5.10	0.30	0.27
β_1	2	1.69	0.75	0.57	1.86	0.75	0.28	2.10	0.57	0.18
λ_1	0.1	0.12	0.13	0.00	0.12	0.12	0.00	0.11	0.10	0.00
λ_2	0.5	0.50	0.15	0.02	0.57	0.14	0.01	0.48	0.14	0.02
γ	0.5	0.50	0.17	0.03	0.49	0.12	0.04	0.51	0.09	0.01
α	0.05	0.03	0.04	0.00	0.04	0.05	0.00	0.04	0.04	0.00
σ	2	2.06	0.13	0.01	2.02	0.06	0.00	2.02	0.06	0.00
D_{11}	2	2.87	0.92	0.62	2.59	0.73	0.53	2.27	0.80	0.22
D_{22}	2	2.03	0.42	0.16	2.21	0.46	0.23	2.10	0.43	0.05
D_{33}	2	2.10	0.37	0.17	0.34	0.50	0.34	2.22	0.59	0.10
D_{44}	4	5.24	1.82	0.76	4.32	0.74	0.60	4.24	0.63	0.18

decrease gradually when the sample size increases. In addition to this, we compared the parameter estimates for different censoring rates (20% and 40%) for a sample size of 500 (Appendix A.4). The result shows that when the sample size is large the censoring rate

has little influence on the estimates.

3.4.2 Simulation study 2

3.4.2.1 Data description

We performed a simulation study on the proportional hazard model having a Gompertz distribution at baseline and non-linear subject-specific trajectory. In particular, the model was of the form

$$h_i(t) = h_0(t)exp(\gamma x_i + \alpha(m_i(t))) = \lambda_1 exp(\lambda_2 t)exp\left\{\gamma x_i + \alpha m_i(t)\right\}, \tag{3.4.5}$$

where $h_0(t)$ is the hazard function at baseline having Gompertz distribution, x_i is baseline covariate and $m_i(t)$ denotes the true and unobserved value of the longitudinal at time t. The observed longitudinal value at time point t for the i^{th} subject had the non-linear form

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= 5\log(1+t) + b_{i1}t + b_{i0} + \varepsilon_i(t), \end{aligned} \tag{3.4.6}$$

where $\varepsilon_i(t) \sim N(0, \sigma^2)$. In the model of (3.4.6), the mean longitudinal response of the population was assumed to have a non-linear logarithmic curve. Different subjects were assumed to have different intercepts and slopes. In particular, we assumed that $b_i = (b_{i0}, b_{i1})^T$ had a bivariate normal distribution with mean $\mu = (3, 2)$ and covariance matrix $D = Diag(1, 1)$. The true values of the other parameters in the model were $\lambda_1 = 0.01$, $\lambda_2 = 0.1$, $\gamma = 0.5$, $\alpha = 0.2$, $\sigma = 2$ respectively. In addition, the censoring mechanism was assumed to be exponentially distributed with parameter $\lambda = 0.25$.

Based on the model in (3.4.5) and simulation study 1, we simulated survival times T_i for 500 subjects with a 35% censoring rate. In particular, the end time for the study was 5 months. All subjects alive at the end of the study (i.e time 5) were assumed to be censored. This design reflected many clinical studies in real life (Burton et al., 2006). In this sample, there were 329 uncensored subjects and 1387 observations for 500 subjects. For each subject, 1-5 longitudinal measurements were recorded. On average, there were 3 longitudinal measurements per subject. In Figure 3.4, the Kaplan-Meier estimate for the survival curve is presented for the simulated data with 95% pointwise CIs in the left panel.

Moreover, the subject-specific longitudinal profiles for six selected subjects is drawn in the right panel. It can be seen that some of the subjects in this dataset showed non-linear evolutions in their longitudinal values. Each subject had its own trajectory.

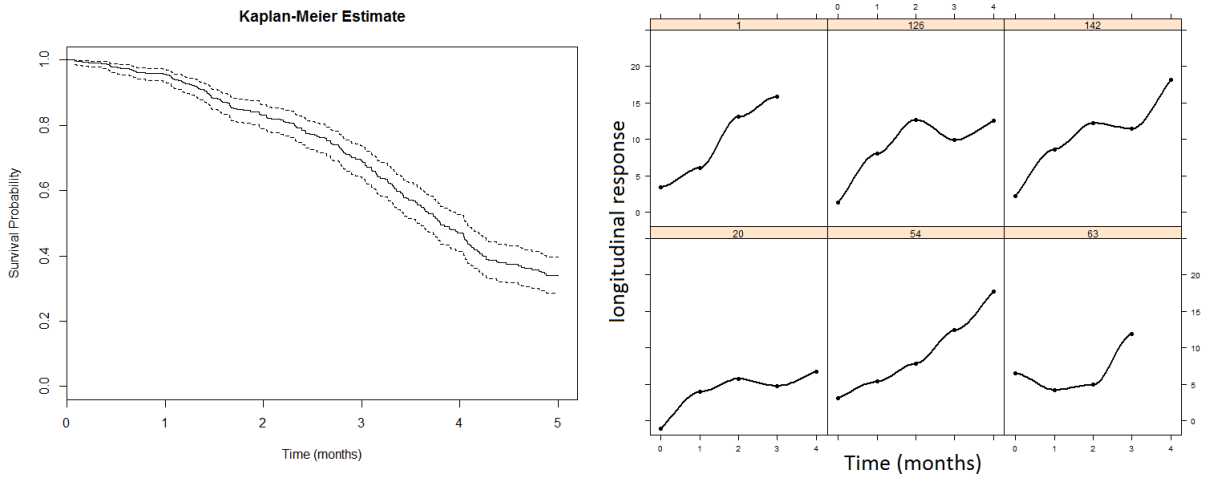

Figure 3.4: Kaplan-Meier estimate of the survival function of the simulated data of (3.4.5) (left panel). Longitudinal trajectories for the six randomly selected subjects of (3.4.6) (right panel).

3.4.2.2 Parameter estimation

We used Model 1 in (3.4.1) and in (3.4.2) to handle the non-linear longitudinal trajectory in the simulated data in (3.4.5). In this model, we put three internal knots at 25%, 50% and 75% respectively of the follow up time. Then, the ECM algorithm as explained in Section 3 was implemented once again to estimate all parameters in the model.

The results for parameter estimation are presented in Table 3.2. The means, standard deviations and 95% CIs of the parameter estimates are calculated for 30 independent samples. The point estimates for λ_1, λ_2 , γ, α, σ^2 are reasonably close to the true values. Similarly, the 95% CIs include the true values of λ_1, λ_2 , γ, α, σ^2.

Based on the estimated values of parameters, we regenerated the estimated survival time by approximating values of random effects from a linear mixed effects function. The detail of the generation is explained in Appendix A.3. Then, we used the Kaplan-Meier estimate to compare the survival function of the simulated dataset (the black solid line) and the estimated survival function (the dashed line), which are presented in the left panel of Figure 3.5.

Table 3.2: Summary statistics for parameter estimation of the simulated data of the model in (3.4.1) and (3.4.2).

Parameter	True value	Estimate	SD	95% CI
β_0	-	3.399	0.673	[3.158;3.640]
β_1	-	4.330	0.142	[4.280;4.380]
λ_1	0.01	0.013	0.021	[0.007;0.021]
λ_2	0.1	0.083	0.184	[0.017;0.148]
γ	0.5	0.640	0.386	[0.486;0.778]
α	0.2	0.186	0.142	[0.136;0.237]
σ	2	1.993	0.061	[1.971;2.015]
D_{11}	-	0.977	0.190	[0.909;1.044]
D_{22}	-	1.365	0.183	[1.300;1.430]
D_{33}	-	1.976	0.154	[1.921;2.031]
D_{44}	-	1.393	0.196	[1.322;1.463]

In addition, we plotted the smooth and predicted longitudinal profiles for 12 patients chosen randomly in the right panel of Figure 3.5. The dot points are the true observed longitudinal values from the simulated data. The solid lines are the smooth longitudinal profiles of the true observed longitudinal values using the Loess smoother and the dashed lines are the predicted profiles of 12 randomly selected individuals. It is clear that the Kaplan-Meier estimates from the simulated data overlaps the Kaplan-Meier estimates based on the predicted value in the left panel of Figure 3.5. The penalized spline regression model in (3.2.8) is a good fit for subject-specific curves in the right panel of Figure 3.5.

3.4.2.3 Model comparison

We applied Model 1 in (3.4.1) and the standard joint model of Rizopoulos to one set of simulated data. The ECM algorithm, as described in Section 3.3.2, was again implemented to estimate all parameters. The standard Gauss Hermite method with ten quadrature points was used to calculate the integrals in the log-likelihood function. The fitted models were as follows:

Fitted model 1: The penalized spline joint model with the hazard function at baseline

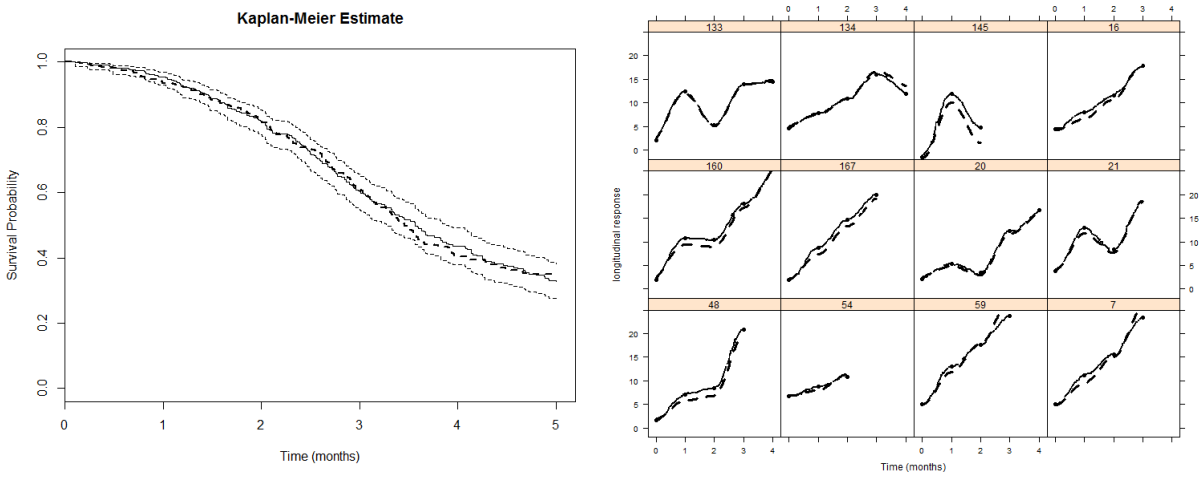

Figure 3.5: Kaplan-Meier estimates of the survival function from simulated failure times (the solid line) with 95% CIs (dot lines), from Model 1 (3.4.1) (the dashed line) (left panel). Observed longitudinal trajectories (the solid line) and predicted longitudinal trajectories (the dashed line) for the twelve randomly selected subjects (right panel).

having Gompertz distribution had the form

$$\begin{cases} \hat{h}_i(t) = 0.0152 \exp(0.1123t) \exp\{0.6319x_i + 0.1824\hat{m}_i(t)\} \\ \hat{m}_i(t) = 3.4396 + 4.3587t + \hat{u}_{i1}(t-\mathcal{K}_1)_+ + \hat{u}_{i2}(t-\mathcal{K}_2)_+ + \hat{u}_{i3}(t-\mathcal{K}_3)_+ + \hat{v}_{i0}\,. \end{cases}$$

Fitted model 2: The standard joint model of Rizopoulos had the form

$$\begin{cases} \hat{h}_i(t) = \hat{h}_0(t) \exp\{0.5622x_i + 0.2750\hat{m}_i(t)\} \\ \hat{m}_i(t) = 3.4432 + 4.2567t + \hat{b}_{0i} + \hat{b}_{1i}t \\ \hat{h}_0(t) = \sum_{q=1}^{7} \hat{\xi}_q I(\nu_{q-1} < t \leq \nu_q)\,. \end{cases}$$

Here $0 = \nu_0 < \nu_1 < ... < \nu_7$ denotes a split of the time scale, with ν_7 being larger than the largest observed time, and ξ_q denotes the value of the baseline hazard in the interval $[\nu_{q-1}, \nu_q)$. Six internal knots were placed at equally spaced percentiles of the observed event times. The values of the baseline hazard in the seven intervals were $\log\hat{\xi}_1 = -4.6227$, $\log\hat{\xi}_2 = -5.2289$, $\log\hat{\xi}_3 = -5.2196$, $\log\xi_4 = -5.5471$, $\log\hat{\xi}_5 = -5.7326$, $\log\hat{\xi}_6 = -6.5182$ and $\log\hat{\xi}_7 = 0.3027$ respectively.

The two most commonly used information criteria are the Akaike's Information Criterion (AIC; Akaike (1974)) and the Bayesian Information Criterion (BIC;Schwarz (1978)). Under these definitions, a model having smaller values of AIC or BIC is considered a better model.

Table 3.3: The maximized log-likelihood, AIC and BIC values for a simulated data.

	Log-Likelihood	AIC	BIC
Model 1	-4169.314	8360.628	8406.238
Model 2	-4179.370	8376.739	8439.959

The maximized values of the log-likelihood function, AIC values and BIC values of the two fitted models are presented in Table 3.3. The results show that the log-likelihood of the penalized spline joint models (fitted model 1) is larger than the log-likelihood value of fitted model 2. This leads to the result that both AIC and BIC values of fitted model 1 are less than the values of fitted model 2. These results confirm that the proposed models can improve the standard joint model and can be effective approaches to handle non-linear longitudinal data.

In summary, simulation studies showed the stability of the algorithm and the goodness of fit of the penalized spline models. From simulation study 1, it was shown that the updating process through the ECM algorithm converged quickly to the true values of the parameters. In addition to this, simulation study 2 showed that the model could well predict the survival function and individual trajectories.

3.4.3 The AIDS data

3.4.3.1 Data description

In the AIDS dataset, there were 467 patients with advanced human imminodeficiency virus infection during antiretroviral treatment who had failed or were intolerant to zidovudine therapy. Patients in the study were randomly assigned to receive either *the didanosine* drug (*ddI*) or *the zalcitabine* drug (*ddC*). CD4 cells are a type of white blood cells made in the spleen, lymph nodes, and thymus gland and are part of the infection-fighting system. CD4 cell counts were recorded at the time of study entry as well as at 2, 6, 12, and 18 months thereafter. Details regarding the design of this study can be found in Abrams et al. (1994). By the end of the study, 188 patients had died, resulting in about 59.7% censoring. There were 1405 longitudinal responses recorded.

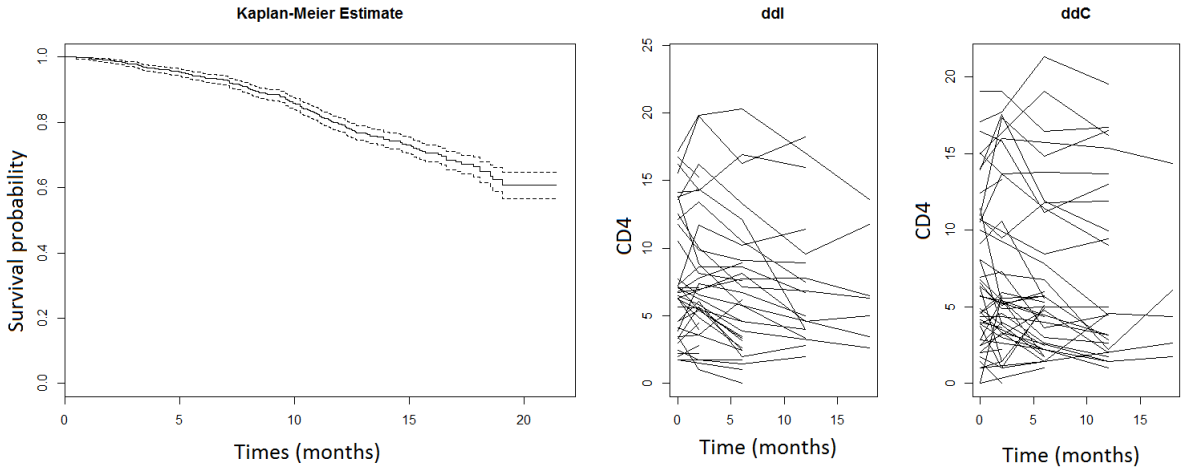

Figure 3.6: Kaplan-Meier estimate of the survival function of the AIDS data (left panel). Longitudinal trajectories for CD4 cell count of the first 100 patients for two groups (right panel).

Previously, Rizopoulos (2012) used his standard joint model for the AIDS data in which the variability between subjects mostly depends on the intercept of the longitudinal submodel. However, the model could not predict observed longitudinal data accurately. When the time unit was changed from month to year in the data, the variability between subjects depended not only on the intercept but also on the slope of the longitudinal submodel. In addition, the longitudinal trajectories plot shows many non-linear curves as depicted in the right panel of Figure 3.6.

Given the nonlinearity, it is appropriate to apply our models, Model 1 and Model 2, to the AIDS data. In particular, we used the two joint models in (3.2.8) and (3.2.9) with three internal knots placed at 25%, 50% and 75% respectively of the observed failure times for the hazard rate at baseline. Then, the ECM algorithm was implemented to estimate all parameters in the two models. A summary of statistics for parameter estimation using Model 1 and Model 2 is presented in Table 3.4.

Following Rizopoulos (2012), in Model 1 and Model 2, the univariate Wald tests were applied for the fixed effects $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ in the longitudinal submodel, the regression coefficient γ and the association parameter α respectively. The results from Table 3.4 show that the point estimates of β_0, β_1, γ, α are all statistically significant for both models at a significance level of 5%.

Table 3.4: Summary statistics for parameter estimation of the AIDS data of Model 1 and Model 2 respectively.

Model 1					Model 2				
Par	Estimate	Std.err	z-value	p-value	Par	Estimate	Std.err	z-value	p-value
β_0	7.87	0.06	127.07	<0.001	β_0	7.81	0.07	114.34	<0.001
β_1	-1.69	0.11	-14.77	<0.001	β_1	-1.62	0.12	-12.99	<0.001
γ	0.22	0.11	2.06	0.039	γ	0.31	0.10	3.03	0.002
α	-0.20	0.01	-15.84	<0.001	α	-0.24	0.01	-18.15	<0.001
λ_1	1.68	0.07			λ_1	1.04	0.11		
λ_2	0.33	0.00			λ_2	1.79	0.23		
σ	2.36	0.36			λ_3	1.38	0.38		
D_{11}	2.18	0.14			λ_4	1.67	0.42		
D_{22}	1.04	0.07			λ_5	2.48	0.66		
D_{33}	0.85	0.06			σ	2.62	0.45		
D_{44}	11.87	0.78			D_{11}	1.02	0.07		
					D_{22}	0.97	0.06		
					D_{33}	0.99	0.07		
					D_{44}	11.40	0.75		

We applied the Kaplan-Meier estimate of the survival function from the observed survival time (the light solid line) and the dotted lines correspond to 95% pointwise CIs in Figure 3.7 (left panel). The predicted survival function from Model 1 is the dashed line and the predicted survival function from Model 2 is the bold solid line. The plots show that Model 2 works very well in this case (Figure 3.7). Moreover, Model 2 is preferred in practice because $h_0(.)$ is usually considered as unspecified in order to avoid the impact of misspecifying the distribution of survival times.

Based on the model for longitudinal regression in (3.4.2), we draw the smooth and predicted longitudinal profiles for 12 patients from the AIDS dataset as depicted in Figure 3.7 (right panel). The dot points are the true observed longitudinal values. The solid lines are the smooth longitudinal profiles using the Loess smoother and the dashed lines are the predicted profiles. Most of the predicted profiles are quite close to the observed ones.

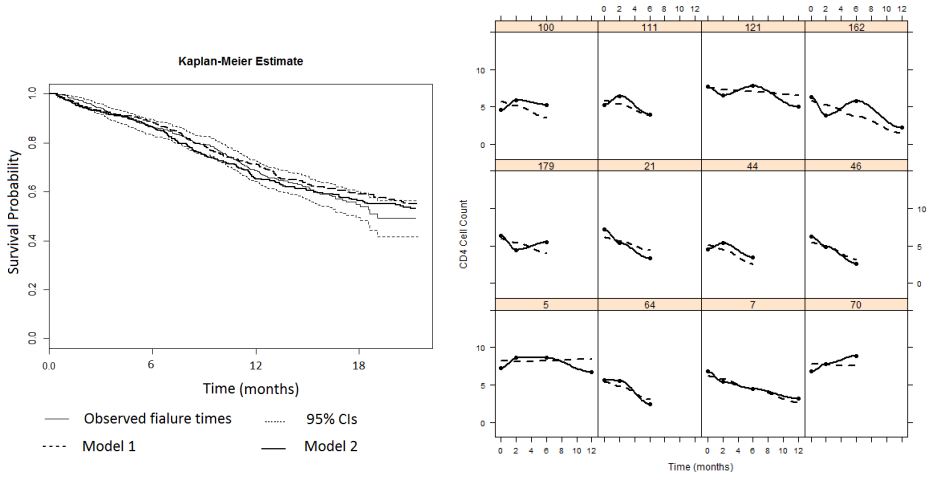

Figure 3.7: Kaplan-Meier estimates of the survival function from observed failure times, from Model 1 and from Model 2 (left panel). Observed longitudinal trajectories (the solid line) and predicted longitudinal trajectories (the dashed line) for the twelve randomly selected patients (right panel).

3.4.3.2 Model comparison

Recall the penalized spline joint Model 1, Model 2 and Rizopoulos's standard joint model used for AIDS data. The ECM algorithm as described in Section 3.2 was implemented to estimate all parameters. The standard Gauss Hermite method with ten quadrature points was used to calculate the integrals in the log-likelihood function in Model 1 and Model 2. For Rizopoulos's joint models (Model 3), the JM package in R language was used to estimate the parameters using the adaptive Gauss Hermite method with five quadrature points to calculate the integrals in the log-likelihood function. The fitted models were as follows:

Fitted model 1: The penalized spline joint model with the hazard function at baseline having the Gompertz distribution in which parameter estimates were taken from the left panel of Table 3.4 had the form

$$\begin{cases} \hat{h}_i(t) = 1.68 \exp(0.33t) \exp\{0.22 drugddI - 0.2\hat{m}_i(t)\} \\ \hat{m}_i(t) = 7.87 - 1.69t + \hat{u}_{i1}(t - \mathcal{K}_1)_+ + \hat{u}_{i2}(t - \mathcal{K}_2)_+ + \hat{u}_{i3}(t - \mathcal{K}_3)_+ + \hat{v}_{i0}\,. \end{cases}$$

Fitted model 2: The penalized spline joint model with the hazard function at baseline having piecewise constant function in which parameter estimates were taken from the

right panel of Table 3.4 had the form

$$\begin{cases} \hat{h}_i(t) = \hat{h}_0(t) \exp\{0.31 drugddI - 0.24 \hat{m}_i(t)\} \\ \hat{m}_i(t) = 7.81 - 1.61t + \hat{u}_{i1}(t - \mathcal{K}_1)_+ + \hat{u}_{i2}(t - \mathcal{K}_2)_+ + \hat{u}_{i3}(t - \mathcal{K}_3)_+ + \hat{v}_{i0} \\ \hat{h}_0(t) = \sum_{q=1}^{5} \hat{\lambda}_q I(\nu_{q-1} < t \leq \nu_q)\,, \end{cases}$$

where the three internal knots in the baseline hazard rate were placed at 25%, 50%, 75% of the observed failure times. The values of the baseline hazard in the five intervals were $\hat{\lambda}_1 = 1.04$, $\hat{\lambda}_2 = 1.79$, $\hat{\lambda}_3 = 1.38$, $\hat{\lambda}_4 = 1.68$ and $\hat{\lambda}_5 = 2.48$.

Fitted model 3: The standard joint model of Rizopoulos (2012) in which parameter estimates were taken from Section 4.2 of Rizopoulos (2012) had the form

$$\begin{cases} \hat{h}_i(t) = \hat{h}_0(t) \exp\{0.33 drugddI - 0.29 \hat{m}_i(t)\} \\ \hat{m}_i(t) = 7.22 - 0.19t + 0.01t * drugddI + \hat{b}_{0i} + \hat{b}_{1i}t \\ \hat{h}_0(t) = \sum_{q=1}^{7} \hat{\xi}_q I(\nu_{q-1} < t \leq \nu_q)\,. \end{cases}$$

Here $0 = \nu_0 < \nu_1 < ... < \nu_7$ denotes a split of the time scale, with ν_7 being larger than the largest observed time, and ξ_q denotes the value of the baseline hazard in the interval $[\nu_{q-1}, \nu_q)$. Six internal knots were placed at equally spaced percentiles of the observed event times. The values of the baseline hazard in the seven intervals were $\log \hat{\xi}_1 = -2.54$, $\log \hat{\xi}_2 = -2.27$, $\log \hat{\xi}_3 = -1.96$, $\log \xi_4 = -2.5$, $\log \hat{\xi}_5 = -2.42$, $\log \hat{\xi}_6 = -2.4$ and $\log \hat{\xi}_7 = -2.42$ respectively.

The maximized value of the likelihood function, AIC value and BIC values of the three models are presented in Table 3.5. The results show that the penalized spline joint models (Model 1 and Model 2) improved the log-likelihood when there are three internal knots in the longitudinal submodel. In a similar way, both AIC and BIC values of fitted models 1 and 2 are significantly lower than the values of fitted model 3. These results confirm that the proposed models can improve the standard joint models. Between fitted models 1 and 2, the fitted model 2 is a preferred model for the AIDS data.

Table 3.5: The maximized log-likelihood, AIC and BIC values for AIDS data.

	Log-Likelihood	AIC	BIC
Model 1	-4236.063	8494.126	8539.736
Model 2	-4201.639	8431.278	8489.327
Model 3	-4328.261	8688.523	8754.864

3.5 Discussion

In this chapter, two joint models using a penalized spline with a truncated polynomial basis have been proposed to model non-linear longitudinal outcomes and time-to-event data. The use of a truncated polynomial basis gives an intuitive way to model non-linear longitudinal outcomes. By adding penalties for the coefficients of the knots and using linear mixed effects models, the smoothing is controlled and the individual curves are specified.

The main findings we may derive from this chapter are at least four-fold: (i) the ECM algorithm provides a reasonably quick convergence algorithm for the proposed models; (ii) the fitted joint models are able to measure the association between the internal time-dependent covariates and the risk of an event; (iii) the two models provide a good prediction for both the longitudinal and survival functions, as indicated by empirical results and (iv) the two models can improve the standard joint models as evidenced in the case study.

The limitations of this study are at least three-fold: (i) the number of internal knots is limited to three due to computational costs; (ii) the polynomial power functions can form an ill-conditioned basis for the models and (iii) the estimation results are sensitive when both random effects and error are not normally distributed.

Based on the limitations, future work will focus on using new methods for approximating the integrals to reduce the computational problems or relaxing the normality assumption. A modified two-stage approach is introduced in Chapter 4 to solve this problem. Furthermore, we will apply a different basis for joint models, that is the penalized B-spline. In

terms of parameter estimation, we are looking at a different approach to estimate the parameters in the models using a Bayesian approach, via MCMC algorithms. This problem will be addressed in Chapter 5.

Chapter 4

A Modified Two-stage Approach for Joint Modelling of Longitudinal and Time-to-event Data

4.1 Introduction

Joint models for longitudinal and time-to-event data are used to link survival outcomes with longitudinal measurements in order to obtain better insight into both processes. In modern survival analysis, the Cox and the extended Cox models have been widely reported in the literature. Proportional hazard models introduced by Cox (1972) have been commonly used for their ability to associate the hazard for an event and covariates. The baseline hazard function in these models is usually considered to be non-parametric. Cox (1975) showed that the partial likelihood can be used to estimate the regression coefficients. A survival package for the Cox models (Therneau, 2014) is also now available in R. However, this approach assumes that the time-dependent covariates are predictable processes and measured without error (Cox 1975;Kalbfleisch and Prentice 2002). With these assumptions, the Cox model only just worked for handling time-independent covariates and external time-dependent covariates. For internal time-dependent covariates, this approach can cause bias and poor coverage properties (Sweeting and Thompson, 2011; Rizopoulos, 2012).

In a joint modelling framework, there are many papers in the literature that deal with internal time-dependent covariates. Standard joint models assume that the hazard rate at a given time of the dropout process is associated with the expected value of the longitudinal responses at the same point in time. This also means that the survival function

depends on the entire longitudinal trajectory. As a result, random effects appear in both longitudinal and survival processes. In order to estimate parameters in these models, the full likelihood approach, which uses shared random effects in the longitudinal and survival submodels, is implemented (Rizopoulos, 2012).

According to Gould et al. (2014) and Sweeting and Thompson (2011), the full likelihood approach is an effective way to investigate the relationship between longitudinal and time-to-event data. However, when the subject-specific trajectories show non-linear curves, longitudinal submodels in the joint model need to be parameterised non-linearly to avoid biases (Rizopoulos, 2011, 2012). This can lead to an increase in the dimension of random effects. As a result, the computational problem becomes significantly more complex when dealing with multi-integrals from the joint log-likelihood function and the survival function respectively. A quick and approximate method for estimation is required to reduce the computational problem and to allow for an easier way to handle extended joint models.

The two-stage approach has been investigated previously by Bycott and Taylor (1998), Self and Pawitan (1992), Tsiatis et al. (1995), and Dafni and Tsiatis (1998). The advantage of this approach is that it can solve the problems of computational complexities in the shared random effects of joint models by using standard mixed-effects and survival software in R in two steps. Firstly, the longitudinal data is fitted. Secondly, the fitted values of the longitudinal process are used as covariates in the joint model. The Cox model is used in the second stage to estimate the survival parameters. This two-stage approach can reduce biases compared with the Cox model approach. However, there still remain more biases compared with the shared random effects approach (Sweeting and Thompson, 2011; Ye et al., 2008). The drawbacks are that the use of the Cox model in the second stage can cause biases, and also that the whole history of the true unobserved longitudinal processes is not used for estimating the survival function (Tsiatis and Davidian, 2004; Sweeting and Thompson, 2011).

The original contributions in this chapter include a new way of estimating parameters for the survival process to reduce the bias in the two-stage approach. In particular, in the first stage, estimated values of parameters in the longitudinal submodel are obtained using linear mixed effects model procedures. As a result, the true unobserved longitudinal data can be evaluated continuously over time, and the whole longitudinal history will be accounted for when estimating the survival function. In the second stage, an approximation

of the expected likelihood function for the complete data is proposed and is used to obtain estimates for the survival process. This approach can improve the previous two-stage approach by eschewing the use of the Cox model in the second stage. The whole history of the true unobserved longitudinal processes is involved for estimating the survival function. Moreover, this estimation is close in spirit to the full likelihood approach in the sense that the expected likelihood function of the complete data is maximized. The computational gains are obvious because the multi-integral calculations of the joint log-likelihood function are avoided. Standard mixed effects software is implemented for the first stage to estimate the parameters of the longitudinal process. The modified two-stage approach is described in detail in Sections 4.2 and 4.3. Extensive simulation studies are presented in Section 4.4 to compare the performance of the ordinary two-stage approach, the modified two-stage approach and the full likelihood approach respectively. As the proposed method is new, R code is written for the second stage to estimate the parameters of the survival process. Another original contribution in this chapter is Section 4.5 which presents the impact of the misspecifying random effects distribution through a simulation study as well as the impact under different censoring rates and different measurement intervals.

This chapter is organized as follows: Section 4.2 describes the ordinary two-stage approach, the full likelihood approach and the modified two-stage approach for joint models. Parameter estimation for the modified two-stage approach is presented in Section 4.3. Simulation studies are conducted in Section 4.4.1 to compare the bias and the accuracy between the three approaches. Moreover, we validate the proposed two-stage approach for penalized spline joint models with high dimension of random effects in Section 4.4.2 and then we apply the method to AIDS data in Section 4.4.3. Random effects misspecification analysis is presented in Section 4.5. Finally, Section 4.6 is the discussion.

4.2 The modified two-stage approach

In this section, we recall the notation from Chapter 3. Suppose that there are n subjects in the longitudinal data and survival data. The observed failure time for the i^{th} subject is denoted as $T_i = \min(T_i^*, C_i)$. Here, T_i^* is the true survival time and C_i denotes the censoring time for the i^{th} subject $(i = 1, ..., n)$. An event indicator is also defined as $\delta_i = I(T_i^* \leq C_i)$ in survival data. The longitudinal data consists of the measurements of

the i^{th} subject $y_{ij} = \{y_i(t_{ij}),\ j = 1, ..., n_i\}$ taken at time points t_{ij}.

The penalized spline joint models for longitudinal data and time to event data in Chapter 3 is postulated from a proportional hazard model to be of the form

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= \lim_{dt \to 0} Pr\left\{t \leq T_i^* < t + dt \mid T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{w}_i\right\}/dt \\ &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}. \end{aligned} \tag{4.2.1}$$

Here $h_0(t)$ is the hazard at baseline and $\boldsymbol{w}_i$ is a vector of baseline covariates (such as treatment indicator, gender of a patient, etc). Furthermore, $\mathcal{M}_i(t) = \{m_i(s),\ 0 \leq s < t\}$ denotes the history of the true unobserved longitudinal process up to time point t.

The longitudinal submodel for the i^{th} subject is given by

$$\begin{cases} y_i(t) & = m_i(t) + \varepsilon_i(t),\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ m_i(t) & = \boldsymbol{X}_i^T(t)\boldsymbol{\beta} + \boldsymbol{X}_i^T(t)\boldsymbol{v}_i + \boldsymbol{Z}_i^T(t)\boldsymbol{u}_i \\ & \boldsymbol{v}_i \sim \mathcal{MVN}(\boldsymbol{0}, \sum),\ \boldsymbol{u}_i \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{D}), \end{cases} \tag{4.2.2}$$

where $\boldsymbol{X}_i$ is the $\left(\sum_{i=1}^{n} n_i \times (p+1)\right)$ matrix of fixed effects; $\boldsymbol{Z}_i$ is the $\left(\sum_{i=1}^{n} n_i \times (p+K+1)n\right)$ matrix of random effects defined as in (3.2.5). The vector $\boldsymbol{\beta}$ is the vector of coefficients whereas the vector of random effects for the i^{th} subject, defined by $\boldsymbol{b}_i$, are considered to be latent variables. We assume that the random effects vector follows a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{G} = Cov(\boldsymbol{b}_i)$. The covariance matrix of random effects $\boldsymbol{b}_i^T = (v_{i0}, ..., v_{ip}, u_{ip1}, ..., u_{ipK})$ is given by

$$\boldsymbol{G} = Cov(\boldsymbol{b}_i) = \begin{bmatrix} \sum & 0 \\ 0 & \boldsymbol{D} \end{bmatrix}.$$

Before presenting the modified two-stage approach for joint models, we first describe in more detail the ordinary two-stage and the full likelihood approaches in the following sections. We have briefly reviewed the two approaches in a general setting in Sections 2.3.2.1 and 2.3.2.2.

4.2.1 Ordinary two-stage approach for joint models

First, we present the existing ordinary two-stage approach proposed by Tsiatis et al. (1995) and Dafni and Tsiatis (1998) to estimate the parameters in the joint models. In this

approach, the longitudinal model is fitted separately first. Hence, the longitudinal fitted values, $\hat{m}_i(t)$, are considered to be covariate in the fitted joint model. Tsiatis et al. (1995) replaced the hazard function $h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w})$ in (4.2.1) by the function $h_i(t \mid \overline{\mathcal{M}}_i(t), \boldsymbol{w})$. Here, $\overline{\mathcal{M}}_i(t)$ denotes the history of the true unobserved longitudinal responses at the grid points up to time t of the i^{th} subject and $\overline{\mathcal{M}}_i(t) = \{m_i(t_1), ..., m(t_j), t_j \leq t\}$. Then the partial likelihood is applied to estimate the survival coefficients and the hazard at baseline. More specifically, in the first stage, the linear mixed effects regressions are fitted to the longitudinal data. The fitted value for the i^{th} subject at time t is $\hat{m}_i(t) = \boldsymbol{X}_i^T(t)\hat{\boldsymbol{\beta}} + \boldsymbol{X}_i^T(t)\hat{\boldsymbol{v}}_i + \boldsymbol{Z}_i^T(t)\hat{\boldsymbol{u}}_i$, where $\boldsymbol{X}_i^T(t)\hat{\boldsymbol{\beta}}$ is the fitted mean response and $\boldsymbol{X}_i^T(t)\hat{\boldsymbol{v}}_i + \boldsymbol{Z}_i^T(t)\hat{\boldsymbol{u}}_i$ is the fitted subject deviation from the mean at time t. In the second stage, the estimates of the survival coefficients, $\boldsymbol{\gamma}$ and α, in (4.2.1) are obtained by maximizing the partial likelihood. The partial likelihood has the form

$$\begin{aligned} PL(\boldsymbol{\gamma}, \alpha) = \sum_{i=1}^{n} \int_0^{\infty} \Bigg\{ & R_i(t) \left\{ \boldsymbol{\gamma}^T \omega_i + \alpha \hat{m}_i(t) \right\} \\ & - \log \left[\sum_j R_i(t) \exp \left\{ \boldsymbol{\gamma}^T \boldsymbol{\omega}_i + \alpha \hat{m}_i(t) \right\} \right] \Bigg\} dN_i(t). \end{aligned} \tag{4.2.3}$$

Here, $N_i(t)$ is the number of events for the i^{th} subject at time t , and $R_i(t)$ is the indicator function of the risk process. If the i^{th} subject is at risk at time t, $R_i(t) = 1$. Otherwise, $R_i(t) = 0$. In addition, using the estimates of the survival coefficients, an estimator for the cumulative hazard function at baseline is given by

$$\hat{H}_0(t) = \sum_{x<t} \left[\frac{dN(x)}{\sum_{i=1}^{n} R_i(t) \exp \left\{ \hat{\boldsymbol{\gamma}}^T \boldsymbol{\omega}_i + \hat{\alpha} \hat{m}_i(t) \right\}} \right]. \tag{4.2.4}$$

Here $dN(t)$ denotes the number of events for the whole sample at time t. This estimator is referred to as the Breslow estimator. One of the advantages of this approach is that it is quick to implement when standard mixed effects software is used for the first stage and survival software is used for the second stage. However, this approach can lead to biases and poor coverage properties (Sweeting and Thompson, 2011). This is mainly due to the fact that survival software implementations are usually based on the assumption that the time-dependent covariates remain constant between examination times. According to Rizopoulos (2012), this assumption is not appropriate and is unrealistic for many internal time-dependent covariates. In addition, if the fitted longitudinal values are used at the

grid points in the partial likelihood function, the survival function does not depend on the whole history of the true unobserved longitudinal data. This can lead to a poor estimate of the survival function $S_i(t)$.

4.2.2 The full likelihood approach for joint models

In the penalized spline joint model (4.2.1), the hazard rate at time t is assumed to depend on the true unobserved longitudinal response at time t. Because of the relationship between the hazard function and survival function, the whole covariate history $\mathcal{M}_i(t)$ affects both the survival and the likelihood functions. This assumption is the important difference between the full likelihood approach and the ordinary two-stage approach. Moreover, the vector of the random effects $\boldsymbol{b}_i$ is assumed to be underlying in both the longitudinal and the survival processes (Rizopoulos, 2012). Under these assumptions, the log-likelihood formulation of the joint models can be written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_i \log p(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}) \\ &= \sum_i \log \int_{\boldsymbol{b}_i} p(T_i, \delta_i, \boldsymbol{y}_i, \boldsymbol{b}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i \\ &= \sum_i \log \int_{\boldsymbol{b}_i} p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i \,. \end{aligned} \tag{4.2.5}$$

The notation $\boldsymbol{\theta} = (\boldsymbol{\theta}_t^T, \boldsymbol{\theta}_y^T, \boldsymbol{\theta}_b^T)^T$ denotes the full parameter vector with $\boldsymbol{\theta}_t = (\gamma^T, \alpha, \theta_{h_0}^T)^T$ denoting the parameter vector for the survival outcomes. $\boldsymbol{\theta}_y = (\boldsymbol{\beta}^T, \sigma_\varepsilon^2)^T$ is the parameter vector for longitudinal outcomes, and $\boldsymbol{\theta}_b = \boldsymbol{G}$. Following Rizopoulos (2012), the observed data score vector for the joint models can be written as:

$$\begin{aligned} S(\boldsymbol{\theta}) &= \sum_i \frac{\partial}{\partial \boldsymbol{\theta}^T} \log \int p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i \\ &= \sum_i \frac{1}{p(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}^T} \int p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i \\ &= \sum_i \frac{1}{p(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta})} \int \frac{\partial}{\partial \boldsymbol{\theta}^T} p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i \\ &= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}^T} \log \{p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\} \\ &\qquad \times \frac{p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)}{p(T_i, \delta_i, \boldsymbol{y}_i; \theta)} d\boldsymbol{b}_i \\ &= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}^T} \log \{p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\} \, p(\boldsymbol{b}_i \mid T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i \,. \end{aligned} \tag{4.2.6}$$

To estimate the parameters in model (4.2.1), the ECM algorithm is implemented as in Section 3.3.2. In particular, to derive the maximum likelihood estimates in (4.2.5), the algorithm obtains the parameter estimates of $\hat{\boldsymbol{\theta}}$ which maximize instead the expected value of the complete data log-likelihood at the i^{th} iteration of

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(it)}) &= \sum_i \int \log\left(p(T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b}_i; \boldsymbol{\theta})\right).p(\boldsymbol{b}_i|T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}^{(it)})d\boldsymbol{b}_i \\ &= \sum_i \int \left(\log p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{y_i}|\boldsymbol{b}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{b}_i; \boldsymbol{\theta})\right) p(\boldsymbol{b}_i|T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}^{(it)})d\boldsymbol{b}_i \,. \end{aligned} \tag{4.2.7}$$

Therefore, it is clear that the complexity of the estimation comes from calculating the multi-integrals in (4.2.7). The multi-integrals with respect to the random effects are in the complete data log-likelihood function and the uni-integrals are in the survival functions. As these integrals do not have closed form solutions, we employed standard Gaussian quadrature rules to approximate the values of the integrals. Obviously, the computational burden increases when the dimension of random effects and the number of quadrature points increase. It is very time-consuming for the algorithm in (3.3.2) to converge when handling non-linear longitudinal data in the JM package of (Rizopoulos, 2010) and described in Chapter 3.

4.2.3 Approximations for parameter estimates and the complete data log-likelihood

In this section, we introduce the following theorem to show the properties of the approximations. These approximations, denoted by $\approx$, will be used in the modified two-stage approach being proposed in Section 4.2.4.

Theorem 1. *Denote* $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_t^T, \hat{\boldsymbol{\theta}}_y^T, \hat{\boldsymbol{\theta}}_b^T)^T$ *as the estimator obtained from the joint model in (4.2.1) and* $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_y^T, \tilde{\boldsymbol{\theta}}_b^T)^T$ *as the estimator obtained from the linear mixed effects model in (4.2.2). As* $\min(n_i) \to \infty$*, the following results hold:*

a) $Pr(\left\|\hat{\boldsymbol{\theta}}_b - \tilde{\boldsymbol{\theta}}_b\right\| > \epsilon) \to 0$

b) $Pr(\left\|\hat{\boldsymbol{\theta}}_y - \tilde{\boldsymbol{\theta}}_y\right\| > \epsilon) \to 0$

c) $S(\boldsymbol{\theta}_t) = \frac{\partial}{\partial \boldsymbol{\theta}_t} l(\boldsymbol{\theta}) \approx \sum_i \frac{\partial}{\partial \boldsymbol{\theta}_t} \log p(T_i, \delta_i, \tilde{\boldsymbol{b}}_i; \boldsymbol{\theta}_t, \boldsymbol{\theta}_y)$ *and*

$$E(\log(p(T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b}_i; \hat{\boldsymbol{\theta}}))) \approx \sum_i \log p(T_i, \delta_i, \tilde{\boldsymbol{b}}_i; \hat{\boldsymbol{\theta}}_t, \tilde{\boldsymbol{\theta}}_y) + \log p(\boldsymbol{y_i}, \tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_y) + \log p(\tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i}) \,,$$

where $\tilde{\boldsymbol{b}}_i = \arg\max_{\boldsymbol{b}} \left\{\log p(\boldsymbol{y}_i, \boldsymbol{b}; \tilde{\boldsymbol{\theta}}_y)\right\}$ *and* $\|.\|$ *denotes the Euclidean vector norm.*

Proof. a) From the Bayesian central limit theorem (Cox and Hinkley, 1979), it follows that as $\min(n_i) \to \infty$,

$$\begin{aligned} p(\boldsymbol{b}_i | T_i, \delta_i, y_i; \boldsymbol{\theta}) &\xrightarrow{p} \mathcal{N}(\tilde{\boldsymbol{b}}_i, \tilde{H_i}^{-1}) \,, \\ p(\boldsymbol{b}_i | y_i; \boldsymbol{\theta}) &\xrightarrow{p} \mathcal{N}(\tilde{\boldsymbol{b}}_i, \tilde{H_i}^{-1}) \,, \end{aligned} \tag{4.2.8}$$

where $\tilde{\boldsymbol{b}}_i = \arg\max_{\boldsymbol{b}} \left\{\log p(\boldsymbol{y}_i, \boldsymbol{b}; \tilde{\boldsymbol{\theta}}_y)\right\}$ and $\tilde{H_i}^{-1} = \frac{-\partial \log p(\boldsymbol{y}_i | \boldsymbol{b}; \tilde{\boldsymbol{\theta}}_y)}{\partial \boldsymbol{b} \partial \boldsymbol{b}^T}|_{\boldsymbol{b}=\tilde{\boldsymbol{b}}_i}$.

By (4.2.6), the score functions with respect to $\boldsymbol{\theta}_b$ from the joint model, $S_{jm}(\boldsymbol{\theta}_b)$, and from the linear mixed effects model, $S_{lmm}(\boldsymbol{\theta}_b)$, can be written as

$$\begin{aligned} S_{jm}(\boldsymbol{\theta}_b) &= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}_b} \log \left\{p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\right\} p(\boldsymbol{b}_i \mid T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i \,, \\ S_{lmm}(\boldsymbol{\theta}_b) &= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}_b} \log \left\{p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\right\} p(\boldsymbol{b}_i \mid \boldsymbol{y}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i \,. \end{aligned}$$

Set

$$S_q(\boldsymbol{\theta}_b) = \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}_b} \log \left\{p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\right\} p(\boldsymbol{b}_i; \tilde{\boldsymbol{b}}_i, \tilde{H_i}^{-1}) d\boldsymbol{b}_i.$$

Thus, as $\min(n_i) \to \infty$,

$$\begin{aligned} S_{jm}(\boldsymbol{\theta}_b) &\xrightarrow{p} S_q(\boldsymbol{\theta}_b), \\ S_{lmm}(\boldsymbol{\theta}_b) &\xrightarrow{p} S_q(\boldsymbol{\theta}_b) \,. \end{aligned} \tag{4.2.9}$$

Furthermore,

$$\|S_{jm}(\boldsymbol{\theta}_b) - S_{lmm}(\boldsymbol{\theta}_b)\| \leq \|S_{jm}(\boldsymbol{\theta}_b) - S_q(\boldsymbol{\theta}_b)\| + \|S_q(\boldsymbol{\theta}_b) - S_{lmm}(\boldsymbol{\theta}_b)\| \,. \tag{4.2.10}$$

By (4.2.10), for any $\epsilon > 0$, the following can be obtained

$$\{\|S_{jm}(\boldsymbol{\theta}_b) - S_{lmm}(\boldsymbol{\theta}_b)\| > \epsilon\} \supset \left\{\|S_{jm}(\boldsymbol{\theta}_b) - S_q(\boldsymbol{\theta}_b)\| > \frac{\epsilon}{2}\right\} \cup \left\{\|S_q(\boldsymbol{\theta}_b) - S_{lmm}(\boldsymbol{\theta}_b)\| > \frac{\epsilon}{2}\right\}.$$

Therefore,

$$\begin{aligned} Pr\left\{\|S_{jm}(\boldsymbol{\theta}_b) - S_{lmm}(\boldsymbol{\theta}_b)\| > \epsilon\right\} \leq\ & Pr\left\{\|S_{jm}(\boldsymbol{\theta}_b) - S_q(\boldsymbol{\theta}_b)\| > \frac{\epsilon}{2}\right\} \\ & + Pr\left\{\|S_q(\boldsymbol{\theta}_b) - S_{lmm}(\boldsymbol{\theta}_b)\| > \frac{\epsilon}{2}\right\}. \end{aligned}$$

Combined with (2.2.2),

$$Pr\left\{||S_{jm}(\boldsymbol{\theta}_b) - S_{lmm}(\boldsymbol{\theta}_b)|| > \epsilon\right\} \to 0\,.$$

Equivalently, the maximum likelihood estimator for $\boldsymbol{\theta}_b$ from the joint model and from the linear mixed effects model converge in probability as $\min(n_i) \to \infty$,

$$Pr(|\hat{\boldsymbol{\theta}}_b - \tilde{\boldsymbol{\theta}}_b| > \epsilon) \to 0\,.$$

b) The convergence is proved in Rizopoulos (2011).

c) By (4.2.5), the score function with respect to $\boldsymbol{\theta}_t$ has the form

$$S(\boldsymbol{\theta}_t) = \frac{\partial}{\partial \boldsymbol{\theta}_t} l(\boldsymbol{\theta}) = \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}_t} \log p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i | T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i.$$

Moreover, by (4.2.9), as $\min(n_i) \to \infty$,

$$S(\boldsymbol{\theta}_t) \xrightarrow{p} \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}_t} \log p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i; \tilde{\boldsymbol{b}}_i, \tilde{H_i}^{-1}) d\boldsymbol{b}_i \approx \sum_i \frac{\partial}{\partial \boldsymbol{\theta}_t} \log p(T_i, \delta_i, \tilde{\boldsymbol{b}}_i; \boldsymbol{\theta}_t, \boldsymbol{\theta}_y).$$

In addition to this, the expected function of the complete data log-likelihood at $\hat{\boldsymbol{\theta}}$ has the form

$$\begin{aligned} E(\log\left(p(T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b}_i; \hat{\boldsymbol{\theta}})\right)) &= \sum_i \int \log\left(p(T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b}_i; \hat{\boldsymbol{\theta}})\right) . p(\boldsymbol{b}_i | T_i, \delta_i, \boldsymbol{y_i}; \hat{\boldsymbol{\theta}}) d\boldsymbol{b}_i \\ &= \sum_i \int \Big(\log p(T_i, \delta_i | \boldsymbol{b}_i; \hat{\boldsymbol{\theta}}) + \log p(\boldsymbol{y_i} | \boldsymbol{b}_i; \hat{\boldsymbol{\theta}}) + \log p(\boldsymbol{b}_i; \hat{\boldsymbol{\theta}})\Big) \\ &\qquad \times p(\boldsymbol{b}_i | T_i, \delta_i, \boldsymbol{y_i}; \hat{\boldsymbol{\theta}}) d\boldsymbol{b}_i\,. \end{aligned} \tag{4.2.11}$$

By (2.2.2), (4.2.11) and by the results from (a) and (b),

$$\begin{aligned} E(\log\left(p(T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b}_i; \hat{\boldsymbol{\theta}})\right)) &\xrightarrow{p} \sum_i \int \Big(\log p(T_i, \delta_i | \boldsymbol{b}_i; \hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\theta}}_y) + \log p(\boldsymbol{y_i} | \boldsymbol{b}_i; \hat{\boldsymbol{\theta}}_y) + \log p(\boldsymbol{b_i}; \hat{\boldsymbol{\theta}}_{b_i})\Big) \\ &\qquad \times p(\boldsymbol{b}_i; \tilde{\boldsymbol{b}}_i, \tilde{H_i}^{-1}) d\boldsymbol{b}_i \\ &\approx \sum_i \log p(T_i, \delta_i, \tilde{\boldsymbol{b}}_i; \hat{\boldsymbol{\theta}}_t, \tilde{\boldsymbol{\theta}}_y) + \log p(\boldsymbol{y_i}, \tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_y) + \log p(\tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i})\,, \end{aligned}$$

as $\min(n_i) \to \infty$. □

4.2.4 A modified two-stage estimation approach

To avoid the weaknesses of the ordinary two-stage approach proposed by Tsiatis et al. (1995), we propose a modified two-stage approach. Here, we use the fitted values of the parameters in the longitudinal process and then approximate the expected function of the complete data log-likelihood. Instead of using the partial likelihood to estimate the regression coefficients of the relative risk model, we apply the approximation method for the full likelihood approach. The one-step Newton-Raphson update is implemented in the second stage.

More specifically, the two stages are as follows:

Stage 1: Fit the linear mixed effects regression for the longitudinal data. In this stage, the coefficient of fixed effects, the variance matrix and the best linear unbiased predictors (BLUPs) of the random effects are obtained. As a result, $\hat{m}_i(t)$ can be evaluated continuously throughout time. This stage can be conducted using linear mixed effects models as described by Laird and Ware (1982) and using software provided by Lindstrom and Bates (1988). In particular, we obtain $\tilde{\boldsymbol{\theta}}_y$ and $\tilde{\boldsymbol{\theta}}_b$ by maximizing the restricted log-likelihood function

$$l(\boldsymbol{\theta}_y, \boldsymbol{\theta}_b) = -\frac{1}{2}\sum_{i=1}^{n} \log\, |\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i| - \frac{1}{2}\sum_{i=1}^{n} \log |\boldsymbol{V}_i| - \frac{1}{2}\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \,, \tag{4.2.12}$$

where $\boldsymbol{V}_i = \left[\begin{array}{cc} \boldsymbol{X}_i & \boldsymbol{Z}_i \end{array} \right] \boldsymbol{G} \left[\begin{array}{cc} \boldsymbol{X}_i & \boldsymbol{Z}_i \end{array} \right]^T + \sigma_\varepsilon^2 \boldsymbol{I}_{n_i}$, $\boldsymbol{I}_{n_i}$ is an $n_i \times n_i$ identity matrix. Given $\tilde{\boldsymbol{\theta}}_y$, the estimated random effects vector, $\tilde{\boldsymbol{b}}_i = [\ \tilde{v}_{i0}, ..., \tilde{v}_{ip} \quad \tilde{u}_{ip1}, ..., \tilde{u}_{ipK}\ ]^T$, is obtained from the formula of the best linear unbiased predictor

$$\tilde{\boldsymbol{b}}_i = E(b_i|y_i) = \boldsymbol{G} \left[\begin{array}{cc} \boldsymbol{X}_i & \boldsymbol{Z}_i \end{array} \right]^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \tag{4.2.13}$$

The fitted longitudinal submodel has the form

$$\hat{y}_i(t) = \hat{m}_i(t) + \varepsilon_i(t) = \boldsymbol{X}_i^T(t)\tilde{\boldsymbol{\beta}} + \left[\begin{array}{cc} \boldsymbol{X}_i(t) & \boldsymbol{Z}_i(t) \end{array} \right]^T \tilde{\boldsymbol{b}}_i + \varepsilon_i(t)\,. \tag{4.2.14}$$

Stage 2: A joint model is fitted using the fitted values of the parameters in stage 1 in the form

$$\begin{aligned} h_i(t) &= h_0(t) \exp(\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha \hat{m}_i(t)) \\ \hat{m}_i(t) &= \boldsymbol{X}_i^T(t)\tilde{\boldsymbol{\beta}} + \left[\begin{array}{cc} \boldsymbol{X}_i(t) & \boldsymbol{Z}_i(t) \end{array} \right]^T \tilde{\boldsymbol{b}}_i\,. \end{aligned} \tag{4.2.15}$$

From Theorem 1, the approximation of the expected function of the complete data log-likelihood at $\hat{\boldsymbol{\theta}}$ is

$$E(\log\left(p(T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b}_i; \hat{\boldsymbol{\theta}})\right)) \approx \sum_i \log p(T_i, \delta_i, \tilde{\boldsymbol{b}}_i; \hat{\boldsymbol{\theta}}_t, \tilde{\boldsymbol{\theta}}_y) + \log p(\boldsymbol{y_i}, \tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_y) + \log p(\tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i}), \tag{4.2.16}$$

as $\min(n_i) \to \infty$. We estimate the parameter for survival process by maximizing the approximation of the expected function of the complete data log-likelihood

$$\sum_i \log p(T_i, \delta_i, \tilde{\boldsymbol{b}}_i; \boldsymbol{\theta}_t, \tilde{\boldsymbol{\theta}}_y) + \log p(\boldsymbol{y_i}, \tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_y) + \log p(\tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i}).$$

Here, the density function of survival time is given by

$$\begin{aligned} p(T_i, \delta_i \mid \tilde{\boldsymbol{b}}_i, \tilde{\boldsymbol{\theta}}_y; \boldsymbol{\theta}_t) &= h(T_i \mid \mathcal{M}_i(T_i), \boldsymbol{w}_i, \tilde{\boldsymbol{\beta}}; \boldsymbol{\theta}_t)^{\delta_i} S(T_i \mid \mathcal{M}_i(T_i), \boldsymbol{w}_i, \tilde{\boldsymbol{\theta}}_y; \boldsymbol{\theta}_t) \\ &= \left[h_0(T_i) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha \hat{m}_i(T_i)\right\}\right]^{\delta_i} \\ &\quad \times \exp\left(-\int_0^{T_i} h_0(s) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha \hat{m}_i(s)\right\} ds\right). \end{aligned} \tag{4.2.17}$$

Moreover, the density function for the longitudinal part given the random effects has the form

$$\begin{aligned} p(\boldsymbol{y_i}|\tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_y) p(\tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i}) &= \prod_j p\left\{y_i(t_{ij})|\tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_y\right\} p(\tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i}) \\ &= \prod_{j=1}^{n_i} \frac{1}{(2\pi \tilde{\sigma_\varepsilon}^2)^{\frac{n_i}{2}}} \exp\left\{-\frac{\| \, y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\tilde{\boldsymbol{\beta}} + \left[\begin{array}{cc} \boldsymbol{X}_i(t_{ij}) & \boldsymbol{Z}_i(t_{ij}) \end{array}\right]^T \tilde{\boldsymbol{b}}_i \, \|^2}{2\tilde{\sigma_\varepsilon}^2}\right\} \\ &\quad \times (2\pi)^{-\frac{q_b}{2}} \det(\tilde{\boldsymbol{G}})^{-1/2} \exp(-\tilde{\boldsymbol{b}}_i^T \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{b}}_i/2). \end{aligned} \tag{4.2.18}$$

4.3 Parameter estimation

In this section, we summarise the proposed two-stage estimation approach to estimate the parameters in the model (4.2.1).

Two-stage maximum likelihood method

Stage 1: Use standard mixed effects software to obtain the estimates of $\tilde{\boldsymbol{\theta}}_y$, $\tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i}$ and $\tilde{\boldsymbol{b}}_i$ respectively.

Stage 2: Estimate the parameters of survival process via one-step of the Newton-Raphson algorithm. In this stage, the steps are as follows:

Step 1: First initialise the parameters of the survival process.

Assume that there are m parameters in the survival vector $\boldsymbol{\theta}_t$ and the starting value of the parameter vector is $\boldsymbol{\theta}_t^{(0)} = (\theta_1^{(0)}, ..., \theta_m^{(0)})$. Based on these initial values and the estimates of $\tilde{\boldsymbol{\theta}}_y$, $\tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i}$ and $\tilde{\boldsymbol{b}}_i$ in Stage 1, calculate

$$\begin{aligned} l(\boldsymbol{\theta}_t^{(0)}) &= \sum_i \log p(T_i, \delta_i, \boldsymbol{y}_i, \tilde{\boldsymbol{b}}_i; \boldsymbol{\theta}_t^{(0)}, \tilde{\boldsymbol{\theta}}_y) \\ &= \sum_i \log p(T_i, \delta_i, \tilde{\boldsymbol{b}}_i; \boldsymbol{\theta}_t^{(0)}, \tilde{\boldsymbol{\theta}}_y) + \log p(\boldsymbol{y_i}, \tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_y) + \log p(\tilde{\boldsymbol{b}}_i; \tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i}) \,. \end{aligned}$$

Step 2: Updating parameters

2.1 Given the current value of parameter vector at the i^{th} iteration $\boldsymbol{\theta}_t^{(it)} = (\theta_1^{(it)}, \theta_2^{(it)}, ..., \theta_m^{(it)})$, calculate the log-likelihood

$$l(\boldsymbol{\theta}_t^{(it)}) = \sum_i \log p(T_i, \delta_i, \boldsymbol{y}_i, \tilde{\boldsymbol{b}}_i; \boldsymbol{\theta}_t^{(it)}, \tilde{\boldsymbol{\theta}}_y, \tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i}) \,.$$

2.2 Propose a new value for the first parameter $\theta_1^{(*)}$. Then, calculate the log-likelihood $l(\boldsymbol{\theta}_t^{(*)})$ where $\boldsymbol{\theta}_t^{(*)} = (\theta_1^{(*)}, \theta_2^{it}, ..., \theta_m^{it})$.

2.3 Set $\boldsymbol{\theta}_t^{(it)} = \boldsymbol{\theta}_t^{(*)}$ if $l(\boldsymbol{\theta}_t^{(*)}) \geq l(\boldsymbol{\theta}_t^{(it)})$, otherwise set $\boldsymbol{\theta}_t^{(it)} = \boldsymbol{\theta}_t^{(it)}$.

2.4 Similarly, based on the value of the parameter vector $\boldsymbol{\theta}_t^{(it)}$, update the new value for the second parameter and continue updating for the last parameter and set $\boldsymbol{\theta}_t^{(it+1)}$.

Step 3: Iterate Step 2 until the algorithm converges numerically.

The commonly used criteria for the convergence of the iterations are

$$l(\boldsymbol{\theta}^{(it+1)}) - l(\boldsymbol{\theta}^{(it)}) < \varepsilon \left(|l(\boldsymbol{\theta}^{(it)}) + \varepsilon| \right) \,,$$

where $\boldsymbol{\theta}^{(it)}$ denotes the parameter values at the i^{th} iteration, the value of ε is chosen at about 10^{-8}.

We employ the one-step Newton-Raphson approach to get the updated $\gamma^{(it+1)}$,$\alpha^{(it+1)}$ and $\boldsymbol{\theta}_{h_0}^{(it+1)}$. In particular,

$$\begin{aligned} S(\boldsymbol{\theta}_t) &= \sum_i \frac{\partial}{\partial \boldsymbol{\theta}_t} \log\left\{p(T_i, \delta_i, \tilde{\boldsymbol{b}}_i; \boldsymbol{\theta}_t, \tilde{\boldsymbol{\theta}}_y, \tilde{\boldsymbol{\theta}}_{\boldsymbol{b}_i})\right\} \\ \hat{\boldsymbol{\theta}}_t^{(it+1)} &= \hat{\boldsymbol{\theta}}_t^{(it)} - \left[\frac{\partial S(\hat{\boldsymbol{\theta}}_t^{(it)})}{\partial \boldsymbol{\theta}_t}\right]^{-1} S(\hat{\boldsymbol{\theta}}_t^{it}) \,. \end{aligned} \tag{4.3.1}$$

The components of the core vector corresponding to $\boldsymbol{\theta}_t$ have the following forms:

$$S(\boldsymbol{\gamma}) = \sum_i \boldsymbol{w}_i \left[\delta_i - \exp(\boldsymbol{\gamma}^T \boldsymbol{w}_i)\right] \int_0^{T_i} h_0(s) \exp\left\{\alpha(\boldsymbol{X}_i^T(t)\tilde{\boldsymbol{\beta}} + \left[\begin{array}{cc} \boldsymbol{X}_i(s) & \boldsymbol{Z}_i(s) \end{array}\right]^T \tilde{\boldsymbol{b}}_i)\right\} ds \,,$$

$$\begin{aligned} S(\alpha) = &\sum_i \delta_i \left\{\boldsymbol{X}_i^T(T_i)\tilde{\boldsymbol{\beta}} + \left[\begin{array}{cc} \boldsymbol{X}_i(T_i) & \boldsymbol{Z}_i(T_i) \end{array}\right]^T \tilde{\boldsymbol{b}}_i)\right\} \\ &- exp(\boldsymbol{\gamma}^T \boldsymbol{w}_i) \frac{\partial}{\partial \alpha} \left[\int_0^{T_i} h_0(s) \exp\left\{\alpha(\boldsymbol{X}_i^T(s)\tilde{\boldsymbol{\beta}} + \left[\begin{array}{cc} \boldsymbol{X}_i(s) & \boldsymbol{Z}_i(s) \end{array}\right]^T \tilde{\boldsymbol{b}}_i)\right\} ds\right] \,, \end{aligned}$$

$$\begin{aligned} S(\boldsymbol{\theta}_{h_0(t)}) = &\sum_i \delta_i \frac{\partial \log h_0(T_i; \boldsymbol{\theta}_{h_0(t)})}{\partial \boldsymbol{\theta}_{h_0(t)}^T} \\ &- exp(\boldsymbol{\gamma}^T \boldsymbol{w}_i) \frac{\partial}{\partial \boldsymbol{\theta}_{h_0(t)}^T} \left[\int_0^{T_i} h_0(s) \exp\left\{\alpha(\boldsymbol{X}_i^T(s)\tilde{\boldsymbol{\beta}} + \left[\begin{array}{cc} \boldsymbol{X}_i(s) & \boldsymbol{Z}_i(s) \end{array}\right]^T \tilde{\boldsymbol{b}}_i)\right\} ds\right] \,. \end{aligned}$$

4.4 Empirical results

In order to compare the performance of the ordinary two-stage approach, the modified two-stage approach and the full likelihood approach, two sets of simulation studies were carried out in this section. In simulation study 1, linear longitudinal and survival data were generated with different censoring rates and measurement occasions. The biases and accuracy of estimates were assessed for the three approaches. In simulation study 2, non-linear longitudinal and survival data were generated. The extended joint model using penalized splines was implemented. The results show that the modified two-stage approach can estimate the survival function well because it can handle a large dimension of random effects. In addition, the extended joint model using penalized splines and the proposed two-stage approach were applied to the AIDS data in a case study.

4.4.1 Simulation study 1

We performed a simulation study on the joint model for linear longitudinal and survival data, which has the form

$$h_i(t) = h_0(t)\exp(\gamma x_i + \alpha(m_i(t))) = \lambda \exp\left\{\gamma x_i + \alpha m_i(t)\right\} . \tag{4.4.1}$$

Here $h_0(t)$ is the hazard function at baseline having an exponential distribution, x_i is the baseline covariate. The form of the true and unobserved value of the longitudinal at time t, $m_i(t)$, is given by

$$m_i(t) = \beta_0 + \beta_1 t + u_{io} + u_{i1}t\,, \tag{4.4.2}$$

where $\boldsymbol{b}_i = (u_{i0}, u_{i1})^T$ is the vector of random effects and is assumed to have a normal distribution.

To simulate the observed survival time T_i of the joint model in (4.4.1), we applied the methods adopted by Bender et al. (2005), Austin (2012) and Crowther and Lambert (2013) to generate the true survival time. In particular, based on the relation between the survival function $S_i(t)$, the cumulative hazard function $H_i(t)$ and the cumulative distribution $F_i(t)$,

$$S_i(t) = \exp(-H_i(t)) = 1 - F_i(t)\,. \tag{4.4.3}$$

Following (4.4.3), we set

$$u = 1 - F_i(T_i)\,, \tag{4.4.4}$$

where u is a random variable with $u \sim \mathcal{U}(0,1)$. The true survival time t is the solution of the equation

$$U = \exp(-H_i(t)) = \exp\left(-\int_0^t h_i(s)ds\right) .$$

We assumed further that the censoring mechanism was exponentially distributed. The observed survival time was the minimum between the censoring time and the true survival time. We generated the survival time T_i for $n = 500$ subjects. The true values of the parameters were $\beta_0 = 5$, $\beta_1 = 2$, $\lambda = 0.1$, $\gamma = 0.5$, $\alpha = 0.05$, $D_{11} = 1$, $D_{12} = 0.5$ and $D_{22} = 1$. Then we generated true longitudinal responses $m_i(t)$ using (4.4.2). The observed longitudinal value at time point t for the i^{th} subject was generated from

$$y_i(t) = m_i(t) + \varepsilon_i(t)\,, \tag{4.4.5}$$

where $\varepsilon_i(t) \sim N(0, \sigma^2)$. The true value of σ was 2.

We considered first the case where the censoring rate was 40% and the longest survival time was 8 years. For measurements taken every six months, 1981 longitudinal responses were recorded. On average, there were four longitudinal measurements per subject. For measurements taken every year, there were 1106 observations for 500 subjects. On average, there were 2 longitudinal measurements per subject. The three approaches were implemented to estimate the parameters in the hazard model (4.4.1). For the ordinary two-stage approach, the linear mixed effects and survival softwares were used for the first stage and the second stage respectively. For the full likelihood approach, Rizopoulos's JM package using the adaptive Gaussian method with five quadrature points was applied. For the modified two-stage approach, R code implementing the algorithm in Section 4.3 was applied.

The bias, standard error (SE) and mean square error (MSE) of the estimates are presented for 6 monthly measurement over 100 simulations (Table 4.1). Because the first stages of the ordinary and proposed two-stage approaches are the same, therefore, the estimates of the parameters in the longitudinal submodel are similar. However, the estimates of the parameters in the survival submodel are significantly different. The biases for the hazard rate at baseline, λ, and the survival coefficients, γ, α, of the proposed two-stage approach reduced significantly, nearly ten times compared to the ordinary two-stage approach. In addition, the mean square errors (MSE) for these parameters of the proposed two-stage approach is also remarkably lower than the ordinary two-stage approach.

The results in Table 4.1 also show that the biases of the estimates for the proposed two-stage and the full likelihood approaches are small and comparable with each other. However, the biases for λ, α, σ^2, D_{12}, and D_{22} of the proposed two-stage approach are slightly smaller than the full likelihood approach. Moreover, MSE and SE of the estimates obtained by the proposed two-stage approach are less compared with the full likelihood approach except for the parameters D_{12} and σ. Note that there is no multi-integral calculation with respect to random effects in the proposed two-stage approach. The average computing time in a single dataset for the proposed two-stage approach was 69.9 s (with standard deviation of 11.65 s). This average computing time was slightly less than the average computing time for the full likelihood approach using the adaptive Gaussian method with ten quadrature points, which was 77.4 s (with standard deviation of 11.4 s).

Similar results were found for yearly measurements (Table 4.2).

Table 4.1: Summary statistics for parameter estimation of the simulated data of the model in (4.4.1) for 6 monthly measurements.

Parameter	True value	The ordinary two-stage approach			The modified two-stage approach		
		Bias	SE	MSE	Bias	SE	MSE
λ	0.2	0.0592	0.0013	0.0037	0.0021	0.0033	0.0011
γ	0.5	0.1750	0.0145	0.0513	0.0075	0.0130	0.0168
α	0.05	0.0090	0.0313	0.0010	0.0006	0.0016	0.0003
β_0	5	0.0207	0.0090	0.0086	0.0224	0.0077	0.0063
β_1	2	0.0559	0.0085	0.0103	0.0482	0.0089	0.0102
σ	2	0.0023	0.0038	0.0015	0.0027	0.0042	0.0017
D_{11}	1	0.0006	0.0115	0.0130	0.0016	0.0110	0.0120
D_{12}	0.5	0.0031	0.0213	0.0449	0.0085	0.0199	0.0394
D_{22}	1	0.0228	0.0088	0.0082	0.0059	0.0098	0.0095

Parameter	True value	The full likelihood approach		
		Bias	SE	MSE
λ	0.2	0.0024	0.0041	0.0017
γ	0.5	0.0067	0.0134	0.0177
α	0.05	0.0034	0.0022	0.0005
β_0	5	0.0123	0.0081	0.0066
β_1	2	0.0129	0.0102	0.0104
σ	2	0.0038	0.0038	0.0014
D_{11}	1	0.0006	0.0217	0.0465
D_{12}	0.5	0.0235	0.0164	0.0271
D_{22}	1	0.0164	0.0195	0.0378

Table 4.2: Summary statistics for parameter estimation of the simulated data of the model in (4.4.1) for yearly measurements.

Parameter	True value	The ordinary two-stage approach			The modified two-stage approach		
		Bias	SE	MSE	Bias	SE	MSE
λ	0.2	0.0183	0.0017	0.0064	0.0006	0.0035	0.0012
γ	0.5	0.1279	0.0510	0.0284	0.0005	0.0170	0.0171
α	0.05	0.0031	0.0035	0.0012	0.0010	0.0017	0.0003
β_0	5	0.0116	0.0101	0.0103	0.0066	0.0080	0.0063
β_1	2	0.0773	0.0109	0.0177	0.0781	0.0113	0.0188
σ	2	0.0039	0.0072	0.0052	0.0011	0.0066	0.0043
D_{11}	1	0.01965	0.0166	0.0275	0.0064	0.0149	0.0220
D_{12}	0.5	0.01663	0.0265	0.0698	0.0305	0.0274	0.0753
D_{22}	1	0.0017	0.0105	0.0109	0.0233	0.0111	0.0128

Parameter	True value	The full likelihood approach		
		Bias	SE	MSE
λ	0.2	0.0103	0.0048	0.0024
γ	0.5	0.0014	0.0110	0.0120
α	0.05	0.0022	0.0024	0.0006
β_0	5	0.0026	0.0094	0.0087
β_1	2	0.0020	0.0095	0.0089
σ	2	0.0163	0.0075	0.0058
D_{11}	1	0.0740	0.0324	0.0948
D_{12}	0.5	0.0015	0.0209	0.0433
D_{22}	1	0.0080	0.0194	0.0375

We now consider the second case of the censoring rate being 20% and the longest survival time was 13 years. In this case, the proposed two-stage approach was applied to estimate parameters in model (4.4.1) with different measurement times. For measurements taken every 6 months, 2340 longitudinal responses were recorded. On average, there were 5 longitudinal measurements per subject. For measurements taken every year, there were 1331 observations for 500 subjects. On average, there were 4 longitudinal measurements per subject. For measurements taken every four years, there were 573 observations for 500 subjects. On average, there was 1 longitudinal measurement per subject.

The bias, SE and MSE of the estimates are presented for every six months, one year and

four years of measurement in Table 4.3. The bias and accuracy of the estimates were lower when longitudinal data was measured every 6 months and every 1 year. When follow-up examinations decreased, the estimates were more biased and less accurate, especially for the parameters of the longitudinal submodel. However, even when the measurement interval was four years, the biases for survival parameters only increased by around 1% compared with measurements taken every six months. These results show the reliability and accuracy of the proposed two-stage.

Table 4.3: Summary statistics for parameter estimation of the simulated data of the model in (4.4.1) for different measurements times.

Parameter	True value	Every 6 months			Every 1 year		
		Bias	SE	MSE	Bias	SE	MSE
λ	0.2	0.0018	0.0034	0.0011	0.0068	0.0034	0.0012
γ	0.5	0.0074	0.0162	0.0261	0.0076	0.0153	0.0232
α	0.05	0.0020	0.0016	0.0003	0.0024	0.0017	0.0003
β_0	5	0.0065	0.0077	0.0059	0.0134	0.0092	0.0086
β_1	2	0.0442	0.0076	0.0076	0.0723	0.0105	0.0162
σ	2	0.0033	0.0036	0.0013	0.0132	0.0067	0.0046
D_{11}	1	0.0050	0.0108	0.0116	0.0105	0.0163	0.0263
D_{12}	0.5	0.0463	0.0163	0.0284	0.0336	0.0281	0.0795
D_{22}	1	0.0155	0.0061	0.0039	0.0156	0.0155	0.0158

Parameter	True value	Every 4 year		
		Bias	SE	MSE
λ	0.2	0.0110	0.0040	0.0017
γ	0.5	0.0092	0.0120	0.0143
α	0.05	0.0038	0.0015	0.0002
β_0	5	0.0117	0.0093	0.0087
β_1	2	0.2186	0.0146	0.0689
σ	2	0.1756	0.0315	0.1291
D_{11}	1	0.1838	0.0425	0.2130
D_{12}	0.5	0.0978	0.0385	0.1567
D_{22}	1	0.0221	0.0135	0.0185

4.4.2 Simulation study 2

The second simulation study was made on a proportional hazard model having a Gompertz distribution at baseline and non-linear subject-specific trajectories (Huong et al., 2016). In this simulated data, we did not limit the end time of the study and longitudinal responses were recorded at the time of study entry as well as at every year thereafter. The joint model has the form

$$h_i(t) = h_0(t)\exp(\gamma x_i + \alpha(m_i(t))) = \lambda_1 \exp(\lambda_2 t) exp(\gamma x_i + \alpha(m_i(t)))\,. \tag{4.4.6}$$

Here $h_0(t)$ is the hazard function at baseline having a Gompertz distribution, x_i is the baseline covariate and $m_i(t)$ denotes the true and unobserved value of the longitudinal at time t. The observed longitudinal value at time point t for the i^{th} subject is

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= 5\log(1+t) + b_{i1}t + b_{i0} + \varepsilon_i(t)\,, \end{aligned} \tag{4.4.7}$$

where $\varepsilon_i(t) \sim N(0, \sigma^2)$. In the model (4.4.7), we simulated the mean longitudinal response in the population having non-linear logarithmic curve. Different subjects were assumed to have different intercepts and slopes. In particular, it was assumed that $b_i = (b_{i0}, b_{i1})^T$ having a bivariate normal distribution with mean $\mu = (3, 2)$ and covariance matrix $D = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The true values of the other parameters put into the model were $\lambda_1 = 0.01$, $\lambda_2 = 0.1$, $\gamma = 0.5$, $\alpha = 0.2$, $\sigma = 2$.

Based on the model in (4.4.6), we simulated the survival time T for 500 subjects in which the end time for the study was not limited. In this sample, there were 229 uncensored subjects comprising 45.8% of the whole sample. There were 1687 observations for 500 subjects. For each subject, 1-10 longitudinal measurements were recorded. On average, there were 4 longitudinal measurements per subject. In Figure 4.1, the Kaplan-Meier estimate for survival curve is presented for the simulated data of the model (4.4.6) with 95% pointwise CIs. Moreover, the subject-specific longitudinal profiles for six randomly selected subjects are drawn in the right panel. It can be seen that some of the subjects in this dataset show non-linear evolutions in their longitudinal values. Each subject has its own trajectory.

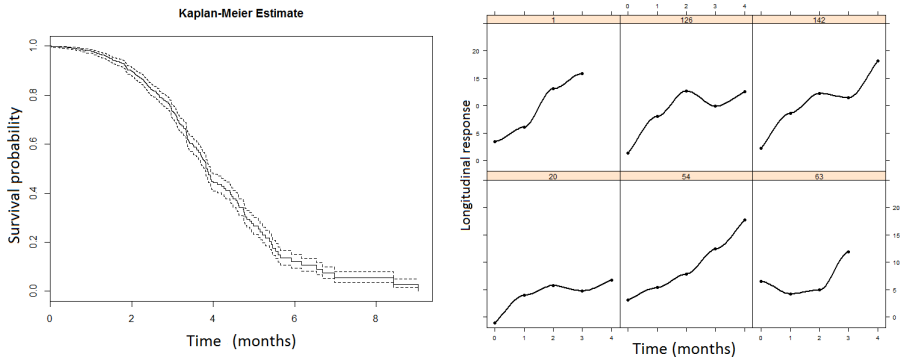

Figure 4.1: Kaplan-Meier estimate of the survival function of the simulated data of (4.4.6) (left panel). Longitudinal trajectories for the six randomly selected subjects of (4.4.7) (right panel).

Penalized spline regression was used to handle the non-linear longitudinal trajectory in the simulated data. The penalized spline submodel has the following form

$$m_i(t) = \beta_0 + \beta_1 t + u_{i0} + u_{i1}(t-\mathcal{K}_1)_+ + u_{i2}(t-\mathcal{K}_2)_+ + u_{i3}(t-\mathcal{K}_3)_+ + ... + u_{ip}(t-\mathcal{K}_K)_+ , \quad (4.4.8)$$

where $\boldsymbol{X}_i^T(t) = [1,\, t]$, $\boldsymbol{Z}_i^T(t) = [1, (t-\mathcal{K}_1)_+, ..., (t-\mathcal{K}_K)_+]$. The set $\{1, t, (t-\mathcal{K}_1)_+, ..., (t-\mathcal{K}_K)_+\}$ is known as the truncated power basis of degree 1. The vector $\boldsymbol{\beta}^T = [\beta_0, \beta_1]$ is called the vector of coefficients and $\mathcal{K}_1, ..., \mathcal{K}_K$ are K fitted knots. We define the vector of random effects for subject i as $\boldsymbol{b}_i^T = (u_{i0},\, u_{i1},\, u_{i2}, ...,\, u_{iK})$. We assumed that the random effects vector follows a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{G} = cov(\boldsymbol{b}_i)$.

Firstly, the *lme* function in R was used to estimate parameters in the longitudinal submodel. Table 4.4 presents the log-likelihood and AIC values for longitudinal process in stage 1 when we put 1 knot, 2 knots, 3 knots, 4 knots and 5 knots into the longitudinal submodel. The results show that the log-likelihood values increase when the number of knots increase. The trend of AIC values is opposite to the log-likelihood values. However, the AIC value for 4 knots is lower than the AIC value for 5 knots, therefore it is the lowest value. According to this result, we should fit the longitudinal submodel with 4 knots at 20%, 40%, 60% and 80% of the follow-up times.

In the first stage, we fitted the joint model in (4.4.7) with 4 knots in the longitudinal submodel. The estimated values from the longitudinal submodel were then put into the

Table 4.4: The log-likelihood and AIC values.

	one knot	two knots	three knots	four knots	five knots
LogLik	-3134.166	-3116.687	-3111.307	-3106.1	-3104.351
AIC	6280.332	6251.375	6248.614	6248.199	6252.702

Table 4.5: Summary statistics for parameter estimation of the simulated data of the model in (4.4.9).

Parameter	True value	Estimate	SD	95% CI
β_0	-	3.3437	0.2226	[3.2820;3.4054]
β_1	-	4.4793	0.2400	[4.4202;4.5533]
λ_1	0.01	0.0215	0.0432	[0.0095;0.0334]
λ_2	0.1	0.0899	0.1055	[0.0607;0.1192]
γ	0.5	0.5391	0.2458	[0.4710;0.5057]
α	0.2	0.1947	0.0896	[0.1698;0.2195]
σ	2	1.9682	0.1848	[1.9169;2.0194]

joint model. The joint model in (4.4.6) was in the form

$$\begin{aligned} h_i(t) &= \lambda_1 \exp(\lambda_2 t) \exp(\gamma x_i + \alpha(\hat{m}_i(t))) \\ \hat{m}_i(t) &= 3.3487 + 4.4703t + \hat{u}_{i0} + \hat{u}_{i1}(t-0)_+ + \hat{u}_{i2}(t-1)_+ + \hat{u}_{i3}(t-2)_+ + \hat{u}_{i4}(t-3)_+ . \end{aligned} \tag{4.4.9}$$

In the second stage, the algorithm in Section 4.3 was implemented to estimate the parameters λ_1, λ_2, γ and α.

The results for parameter estimation are presented in Table 4.5. The estimated mean, SD and 95% CIs of parameter estimates are calculated for 50 independent samples. It can be seen that the point estimates for λ_1, λ_2, γ, α and σ are reasonably close to the true values. Similarly, the 95% CIs include the true values of λ_1, λ_2, γ, α and σ.

Based on the estimated values of parameters, we generated the estimated survival time. Then we used the Kaplan-Meier estimates to compare between the survival function from the simulated dataset (the black solid line) and the estimated survival function from the joint model in (4.4.6) (the dashed line) as presented in the left panel of Figure 4.2. It is clear that the Kaplan-Meier estimates from simulated data overlapped well with the Kaplan-Meier estimates based on the predicted values from the beginning of the study

to the end of the study. In the right panel of Figure 4.4.7, we also draw the smooth and predicted longitudinal profiles for 12 patients chosen randomly. The dot points are the true observed longitudinal values from the simulated data. The solid lines are the smooth longitudinal profiles of the true observed longitudinal values created using the Loess smoother and the dashed lines are the predicted profiles of 12 randomly selected individuals. It can be seen that the penalized spline regression model in (4.4.7) provides a good prediction for the subject-specific curves.

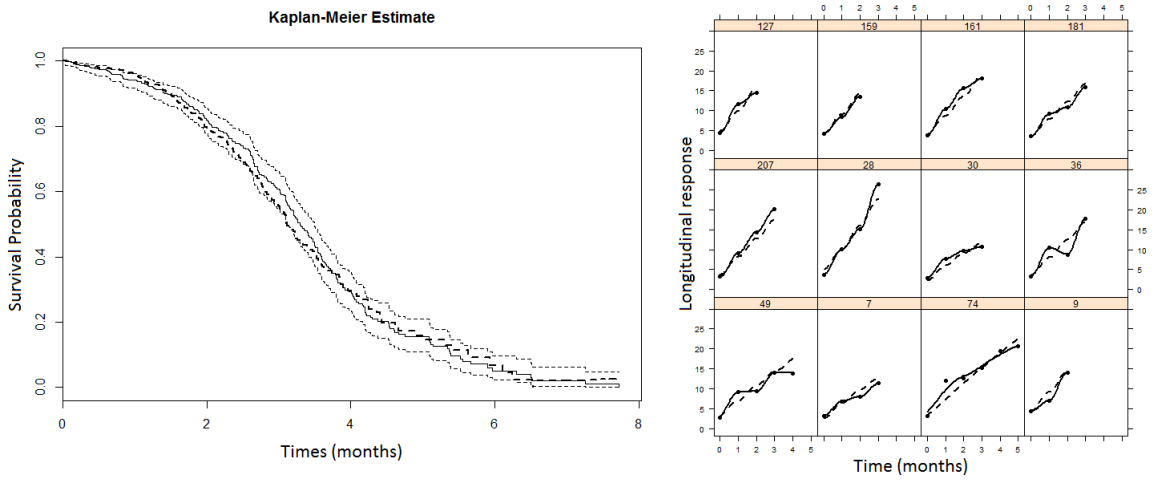

Figure 4.2: Kaplan-Meier estimates of the survival function from simulated failure times (the solid line) with 95% CIs (dot lines), from model in (4.4.9) (the dashed line) (left panel). Observed longitudinal trajectories (the solid line) and predicted longitudinal trajectories (the dashed line) for the twelve randomly selected patients (right panel).

4.4.3 The AIDS data

The penalized spline joint model was applied to the AIDS dataset. The design of this study can be found in Abrams et al. (1994) and the details of this study were presented in Section 3.4.3. In the penalized spline joint model, we put three internal knots in the longitudinal submodel at 25%, 50% and 75% of follow-up time. We assumed that the hazard rate at baseline has a Gompertz distribution. The joint model has the form

$$h_i(t) = h_0(t)exp(\gamma x_i + \alpha(m_i(t))) = \lambda_1 exp(\lambda_2 t)exp(\gamma x_i + \alpha(m_i(t)))\,. \tag{4.4.10}$$

Here the observed longitudinal value at time point t for the i^{th} subject is

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= \hat{\beta}_0 + \hat{\beta}_1 t + \hat{u}_{i0} + \hat{u}_{i1}(t-\mathcal{K}_1)_+ + \hat{u}_{i2}(t-\mathcal{K}_2)_+ + \hat{u}_{i3}(t-\mathcal{K}_3)_+ + \varepsilon_i(t)\,. \end{aligned} \tag{4.4.11}$$

Table 4.6: Summary statistics for parameter estimation of the simulated data of the model in (4.4.10).

Parameter	Estimate	Std.err	z-value	p-value
β_0	2.5080	0.0303	82.8690	<0.0001
β_1	-0.3938	0.0297	-13.2396	<0.0001
λ_1	0.4060	0.0303	13.4164	<0.0001
λ_2	0.5813	0.0907	6.4099	<0.0001
γ	0.2204	0.0502	2.3986	0.01645
α	-0.1955	0.0319	-6.1274	<0.0001
σ	0.3627	-	-	-

From the assumptions of the proposed two-stage approach, the estimates are sensitive to the normal assumptions for random effects and error terms. By the fact that the CD4 cell counts had a distribution skewed to the right, we transform the CD4 cell counts into the square root of the CD4 cell counts. In addition to this, the time unit is changed from months to years in the data. Finally, the algorithm in Section (4.3) is applied to estimate the parameters in the model (4.4.10). The estimated parameters are shown in Table 4.6. The standard errors of the estimates are small and the point estimates are statistically significant at a 5 % significance level.

We draw the Kaplan-Meier estimates of the survival function from the observed survival time (the light solid line) and the dot lines correspond to 95% pointwise CIs in Figure 4.3 (left panel). The predicted survival function from the model in (4.4.10) is the dashed line. In the right panel of Figure 4.3, we also draw the smooth and predicted longitudinal profiles for nine patients chosen randomly. It is shown that the proposed two-stage approach can predict well both the survival function and the subject-specific longitudinal trajectories.

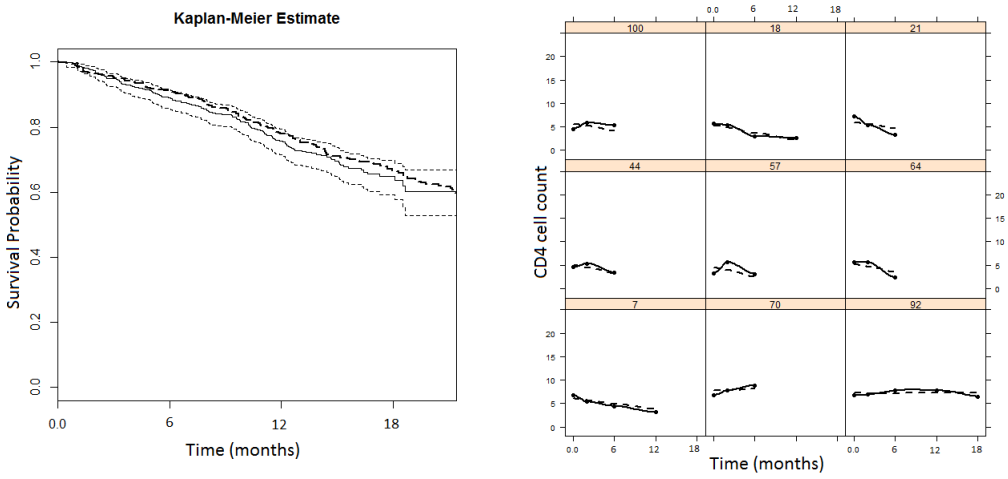

Figure 4.3: Kaplan-Meier estimates of the survival function from observed failure times (the solid line) with 95% CIs (dot lines), from model (4.4.10) (the dashed line) (left panel). Observed longitudinal trajectories (the solid line) and predicted longitudinal trajectories (the dashed line) for the nine randomly selected patients (right panel).

4.5 Random effects misspecification analysis

The joint modelling framework is based on the assumption that the random effects have multinormal distribution with mean zero and covariance matrix G. However, the validity of this assumption is misspecified in practice. The estimation for the parameters in the joint models depends on this assumption. Therefore, this misspecification of the random effects can affect the parameter estimates in the joint models using the full likelihood approach.

In this section, we investigate the impact of misspecifying the random effects distribution through a simulation study. In particular, two mixture distributions are considered for the random effects. The first distribution is the bimodal mixture distribution and the second distribution is the unimodal skewed mixture distribution. In addition, we also consider the impact under different censoring rates and different measurement intervals.

4.5.1 Study set-up

We generated the longitudinal and survival data from the join model

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\} \\ &= \lambda \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\} , \end{aligned} \tag{4.5.1}$$

where $h_0(t)$ has exponential distribution and the longitudinal submodel has the form

$$m_i(t) = \beta_0 + \beta_1 t + u_{i0} + u_{i1} t \, . \tag{4.5.2}$$

In the first case, a bimodal mixture distribution is considered having the form

$$0.4 \times \mathcal{N}\left\{(-2,-2)^T, \boldsymbol{D}\right\} + 0.6 \times \mathcal{N}\left\{(1.333, 1.333)^T, \boldsymbol{D}\right\} , \tag{4.5.3}$$

where $\boldsymbol{D} = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. In the second case, a unimodal skewed mixture distribution is considered having the form

$$0.7 \times \mathcal{N}\left\{(-1,1)^T, \boldsymbol{D}\right\} + 0.3 \times \mathcal{N}\left\{(2.333, -2.333)^T, \boldsymbol{D}\right\} , \tag{4.5.4}$$

where $\boldsymbol{D} = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. The two dimensional pictures for the random effects distribution are presented in Figures 4.4 and 4.5.

The steps for generating the data are as in Section 4.1. The true values for the parameters in the joint model are $\lambda = 0.2$, $\gamma = 0.5$, $\alpha = 0.05$, $\beta_0 = 5$, $\beta_1 = 2$. The observed longitudinal value for the i^{th} subject at time point t has the form

$$y_i(t) = m_i(t) + \varepsilon_i(t) \, ,$$

where the measurement error is assumed to have normal distribution with mean is 0 and variance $\sigma = 2$.

Based on the model in (4.5.1), we simulated the survival time for 500 subjects. The censoring mechnism had an exponential distribution. Here, we considered two cases. When the censoring rate was 40%, the longitudinal measurement were taken every one year and $\max_i(n_i) = 11$. When the censoring rate was 40%, the longitudinal measurement were taken every four years and $\max_i(n_i) = 3$.

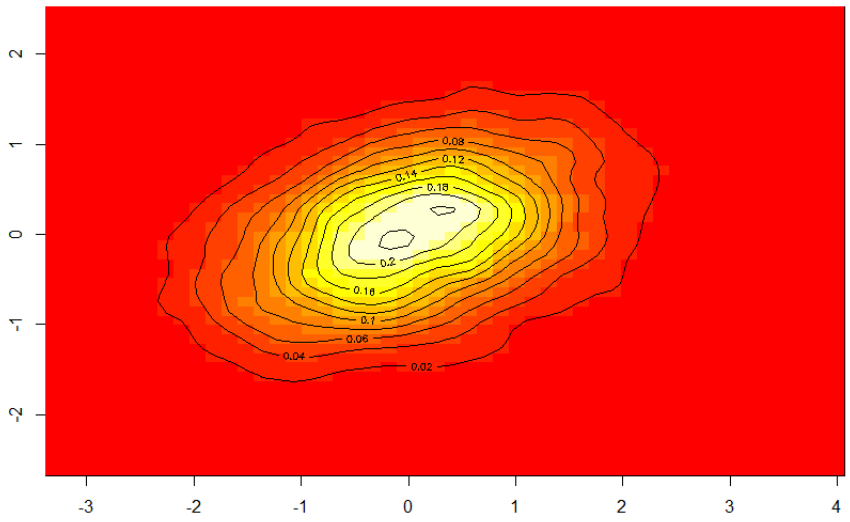

Figure 4.4: The contour plot for the bimodal mixture distribution for the random effects in (4.5.3).

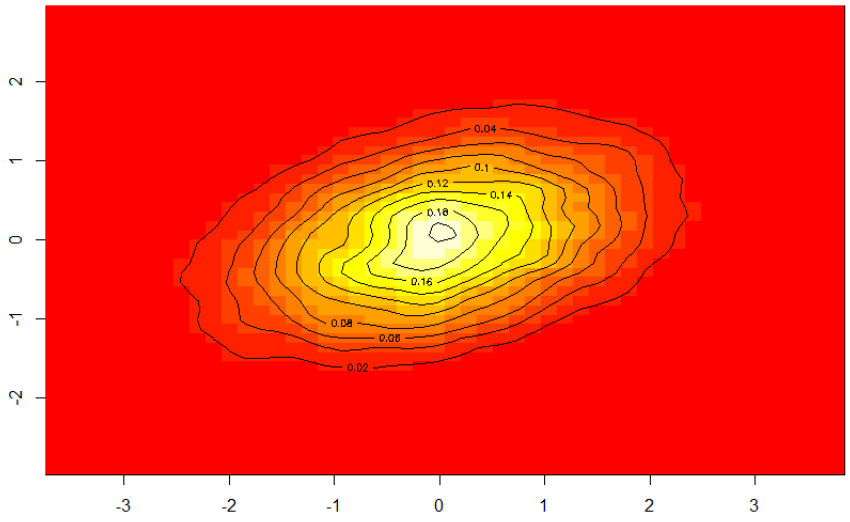

Figure 4.5: The contour plot for the unimodal skewed mixture distribution for the random effects in (4.5.4).

4.5.2 Results

We generated 50 independent datasets from Model 1 in (4.5.1) for different random effects distributions and measurement intervals. The means of the estimates, SD and 95% CIs using the full likelihood approach are presented in Table 4.7. In this table, the 40% censoring rate, the large n_i case and the small n_i case were considered. The estimates when the random effects vector has a bimodal mixture distribution are presented in the upper half and the estimates when the random effects vector has a unimodal skewed mixture distribution are presented in the lower half.

In the upper half, we can see that when $\max_i(n_i)$ is large, the impact of misspecifying the random effects distribution is very minor for all of the parameters. The estimates for the survival parameters, λ, γ and α, are slightly different for both the large n_i case and the small n_i case. When $\max_i(n_i)$ is large, the estimates for the survival parameters are better and it is very close between the bimodal mixture random effects and Gausian random effects.

In the other hand, the longitudinal paramters, β_0, β_1 and σ_ε are affected when $\max_i(n_i)$ is small. The bias and the accuracy of these parameters for the small n_i case are greater than in the large n_i case, especially for the error measurement σ_ε. In particular, the bias and variation for the measurement error increase when $\max_i(n_i)$ decreases. In the lower half, the results are the same with the random effects having a unimodal skewed mixture distribution for the the large n_i case and the small n_i case. These results again show that when $\max_i(n_i)$ is large the joint modelling can reduce the impact of the misspecifying random effects distribution.

Table 4.7: Summary statistics for parameter estimation of the simulated data of the model in (4.5.1) for 40% censoring rate and different measurement intervals. The upper half contains the results for the random effects having a bimodal mixture distribution, whereas the lower half contains the results for the random effects having a unimodal skewed mixture distribution.

Bimodal mixture		Every 1 year			Every 4 years		
Parameter	True	Mean	SD	95% CI	Mean	SD	95% CI
λ	0.2	0.226	0.041	[0.164;0.326]	0.239	0.047	[0.164;0.327]
γ	0.5	0.488	0.114	[0.260;0.716]	0.460	0.119	[0.237;0.703]
α	0.05	0.035	0.019	[0.001;0.072]	0.035	0.014	[0.002;0.073]
β_0	5	5.074	0.158	[4.761;5.389]	4.896	0.291	[4.762;5.416]
β_1	2	1.847	0.201	[1.564;2.144]	1.757	0.278	[1.546;2.298]
σ	2	2.171	0.210	[1.885;2.261]	2.232	0.408	[1.964;2.371]
Unimodal skewed mixture		Every 1 year			Every 4 years		
Parameter	True	Mean	SD	95% CI	Mean	SD	95% CI
λ	0.2	0.221	0.039	[0.162;0.312]	0.242	0.042	[0.170;0.331]
γ	0.5	0.486	0.111	[0.241;0.694]	0.467	0.113	[0.260;0.711]
α	0.05	0.036	0.016	[0.005;0.066]	0.031	0.018	[0.002;0.070]
β_0	5	5.009	0.159	[4.775;5.244]	5.092	0.160	[4.771;5.409]
β_1	2	1.886	0.201	[1.534;2.110]	1.816	0.243	[1.501;2.269]
σ	2	2.093	0.264	[1.858;2.184]	2.231	0.471	[1.812;2.448]

4.6 Discussion

In this chapter, a modified two-stage approach has been proposed to estimate parameters in the joint models for longitudinal and survival data. This approach can reduce the computational challenges by avoiding the calculation for multi-integrations in the full likelihood approach. This allows the application of extended longitudinal submodels with a high dimension of random effects in joint models to handle non-linear longitudinal data. Moreover, in our proposed two-stage model, survival parameters are estimated by maximizing the approximation of the fully log-likelihood function of joint models. By doing this, the proposed two-stage approach improves on weaknesses of the existing two-stage approach and reduces biases. In addition, in simulation studies and a case study,

this approach performs very well and is comparable to the full likelihood approach.

Simulation study 1 shows that the proposed two-stage approach has reduced the biases and improved the accuracy quite significantly for the parameters in the survival submodels compared to the ordinary two-stage approach. This simulation study also shows comparable results with the proposed two-stage approach and the full likelihood approach for the bias and the accuracy respectively. Note that when the dimension of random effects increases, the running time for the two-stage approach is noticeably less than for the full likelihood approach. This is because of the avoidance of the multi-integral calculation by using the LMEs and BLUPs. Moreover, the results also show that the proposed two-stage approach is reliable for estimating the survival parameters when we change the time interval of measurements.

Simulation study 2 and the case study show that the proposed two-stage approach can allow the application of the extended joint models with a high dimension of random effects. The better the longitudinal submodel that can be fitted, the better the model can predict the survival functions and subject-specific trajectories. In addition to these findings, the effect of misspecification of the random effects distribution can be reduced when the number of measurements from the longitudinal process increases.

There are at least three limitations to this approach. Firstly, the prediction of random effects using BLUPs depends critically on the assumption of normally distributed random effects and error terms. Secondly, the uncertainty in the longitudinal submodel estimations does not affect the estimation of the survival submodel. In addition, the highly informative dropout can cause biases on estimating parameters in the longitudinal submodel. To overcome these problems, transformations for longitudinal covariates need to be considered to satisfy the normal assumption. The variability of the estimates from the first stage can properly be taken into account by using the Monte Carlo method proposed in Chapter 5.

Chapter 5

Parameter Estimation for The Penalized Spline Joint Models: A Bayesian Approach

5.1 Introduction

In classical analysis, estimates are usually based on the likelihood function. To apply the full likelihood approach for estimating parameters in the joint models as presented in Chapter 3, we have to deal with multi-integrals with respect to the random effects. This can lead to computational complexity and unstable estimations (Rizopoulos, 2012). In this chapter, we will apply a fully Bayesian approach for the penalized spline joint models. In this approach, the asymptotic approximations for the integral solution are not needed (Ibrahim et al., 2005; Geman and Geman, 1984; Gelman et al., 1995). Instead, parameters in the joint models are sampled through target posterior distributions. By doing this, the uncertainties of parameters can be fully inferred through their marginal posterior densities. In addition, this approach can make good use of historical data embedded in their priors (Gould et al., 2014).

Recently, there are different Bayesian statistical methods for joint models. Faucett and Thomas (1996); Wang and Taylor (2001) implemented a Bayesian method for the joint model having a mixed-effects longitudinal submodel and piecewise-constant hazard rate at baseline. Faucett and Thomas (1996) introduced non-informative priors for all the parameters. In particular, improper uniform priors are specified for the fixed-effects and random effects and improper priors are specified for the remainder of the parameters. The estimates for all unknown parameters are obtained using Gibbs Sampling. Wang

and Taylor (2001) also applied a random effects model with univariate distribution and include an integrated Ornstein-Uhlenbeck longitudinal submodel. This method adds more flexibility to the subject-specific curves, but it can cause an increase in the number of parameters (Tsiatis and Davidian, 2004; Gould et al., 2014).

To relax the assumption for the random effects in (2.3.3), Brown and Ibrahim (2003) introduced a mixture of Dirichlet process models (DPM) for the joint models. They used a quadrature form for the longitudinal submodel. However, to provide a good fit to the non-linear longitudinal data, Brown et al. (2005) and Rizopoulos (2014) applied a Bayesian approach to joint models having a B-spline longitudinal submodel. Brown et al. (2005) chose the proper prior distributions that conjugate to the likelihood. Then, the GS algorithm is applied to obtain samples from the posterior distribution. Rizopoulos (2014) proposed a Bayesian approach for the joint models having a generalized linear mixed effects model. The MH and slice sampling algorithms were used to sample parameters.

In this chapter, a fully Bayesian approach is proposed for the penalized spline joint models introduced in Chapter 3. Firstly, we take full advantage of the ordinary two-stage approach in order to define the prior distributions for the parameters in the joint model (Rizopoulos, 2014). To implement a Bayesian approach, the joint posterior distribution of the joint model is derived using the proposed prior distributions. A set of MCMC algorithms is then proposed to sample parameters from the conditional posterior distributions (Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984). In particular, the random walk MH algorithms are applied to sample for survival coefficients and longitudinal coefficients and independent MH algorithms are implemented for sampling random effects. The GS algorithms are used to sample for the measurement errors and the random effects precision matrix. Before presenting the statistical inferences, we also implement the Gelman and Rubin and the Geweke diagnostics to check for the convergence of the chains.

The prior distributions that are chosen for unknown parameters can have an impact on inferences (Gelman et al., 1995; Wakefield, 2013). Therefore, a prior sensitivity analysis needs to be conducted to validate statistical inferences using Bayesian approach. In the joint modelling framework, the hazard rate at baseline is an unspecified part. There is minimal information for parameters in the hazard function at baseline. In addition, the association parameter between longitudinal data and survival data is the most important

parameter to evaluate the impact of subject specific information on its survival time. In this thesis, we conduct a prior sensitivity analysis on these key parameters.

In summary, the original contributions in this chapter include:

(i) The detail derivation of joint and conditional posterior distributions of the parameters in the proposed model in Section 5.3;

(ii) The MCMC main algorithm in Section 5.4 which consists of the random walk MH algorithms for survival coefficients and longitudinal coefficients; the independent MH algorithms for sampling random effects and the GS algorithms for the measurement errors and the random effects precision matrix;

(iii) Extensive simulation studies in Section 5.5 to validate the proposed MCMC algorithms in (ii);

(iv) A prior sensitivity analysis for the parameter of hazard at baseline and the association parameter between longitudinal data and survival data of the proposed model in Section 5.6.

This chapter is organized as follows. Section 5.2 describes the penalized spline joint model through a three-stage hierarchical model. In this section, two specific joint models are introduced. Prior distributions, likelihood functions and the joint posterior distribution are detailed in Sections 5.3. In Section 5.4, a set of MCMC algorithms is introduced. We then apply the proposed algorithms to extensive simulations studies in Section 5.5. The prior sensitivity analysis is presented in Section 5.6 followed by a case study in Section 5.7. The conclusion is discussed in Section 5.8.

5.2 A three-stage hierarchical for the penalized spline joint models

First recall the notation and models introduced in Section 3.2. There are n subjects in the longitudinal data and survival data. The observed failure time for the i^{th} subject is denoted as $T_i = \min(T_i^*, C_i)$. Here, T_i^* is the true survival time and C_i denotes the censoring time for the i^{th} subject $(i = 1, ..., n)$. An event indicator is defined as $\delta_i =$

$I(T_i^* \leq C_i)$ in survival data. The longitudinal data consists of the measurements of the i^{th} subject $y_{ij} = \{y_i(t_{ij}),\ j = 1, ..., n_i\}$ taken at time points t_{ij}. Therefore, the observed data for the joint models consists of a count of the number of (T_i, δ_i, y_i), $i = 1, .., n$.

Using the penalized spline joint models in Chapter 3, the joint model for longitudinal data and time to event data is postulated from a proportional hazard model of the form

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= \lim_{dt \to 0} Pr\left\{t \leq T_i^* < t + dt \mid T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{w}_i\right\}/dt \\ &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}. \end{aligned} \tag{5.2.1}$$

Here, $h_0(t)$ is the hazard at baseline and $\boldsymbol{w}_i$ is a vector of baseline covariates. Furthermore, $\mathcal{M}_i(t) = \{m_i(s),\ 0 \leq s < t\}$ denotes the history of the true unobserved longitudinal process up to time t. The longitudinal submodel can be written as

$$\begin{cases} y_i(t) & = m_i(t) + \varepsilon_i(t),\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ m_i(t) & = \boldsymbol{X}_i^T(t)\boldsymbol{\beta} + \boldsymbol{X}_i^T(t)\boldsymbol{v}_i + \boldsymbol{Z}_i^T(t)\boldsymbol{u}_i \\ & \boldsymbol{v}_i \sim \mathcal{MVN}(\boldsymbol{0}, \sum),\ \boldsymbol{u}_i \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{D}). \end{cases} \tag{5.2.2}$$

Here, we set $\boldsymbol{b}_i = (\boldsymbol{v}_i^T, \boldsymbol{u}_i^T)^T$ which is the random effects vector of the joint model. We assume that the random effects vector follows a normal distribution with mean 0 and covariance matrix $\boldsymbol{G}$, $\boldsymbol{b}_i \sim \mathcal{MVN}(0, \boldsymbol{G})$. Here,

$$\boldsymbol{G} = Cov(\boldsymbol{b}_i) = \begin{bmatrix} \sum & 0 \\ 0 & \boldsymbol{D} \end{bmatrix}.$$

Given the random effects, the longitudinal process is assumed to be independent with the event time process. Moreover, the longitudinal responses of each subject are assumed independent. In particular, the joint likelihood function of observed survival times and observed longitudinal outcomes is shown to be

$$\begin{aligned} p(T_i, \delta_i, \boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\theta}) &= p(T_i, \delta_i | \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\theta}), \\ p(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\theta}) &= \prod_{j=1}^{n_i} p(y_{ij} | \boldsymbol{b}_i, \boldsymbol{\theta}), \end{aligned} \tag{5.2.3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_t^T, \boldsymbol{\theta}_y^T, \boldsymbol{\theta}_b^T)^T$ denotes the full parameter vector with $\boldsymbol{\theta}_t = (\boldsymbol{\gamma}^T, \alpha, \boldsymbol{\theta}_{h_0}^T)^T$ denoting the parameter vector for the survival outcomes. Furthermore, $\boldsymbol{\theta}_y = (\boldsymbol{\beta}^T, \sigma_\varepsilon^2)^T$ is the parameter vector for longitudinal outcomes and $\boldsymbol{\theta}_b = \boldsymbol{G}$ is the vector of the variance matrix of random effects.

It is important to take into account the domain of the full parameter vector in order to determine prior and posterior distributions in a Bayesian setting. The elements of the parameter vector of the hazard at baseline, $\boldsymbol{\theta}_{h_0}$, are positive real values. In addition, the regression coefficients in the survival submodel, $\boldsymbol{\gamma}$, α and the regression coefficients in the longitudinal submodel, $\boldsymbol{\beta}$ are real values. The variance for the error measurement, σ_ε^2, is always positive while the variance matrix of random effects, $\boldsymbol{G}$, has positive values on the main diagonal and real numbers elsewhere.

The function $h_0(.)$ is the unknown part in the joint model. Thus, to specify the model in (5.2.1), we need to determine the form of the function $h_0(.)$. In this chapter, standard options with known parametric distributions are used for the risk function at baseline (**?**). The exponential and Gompertz distributions are chosen.

There are two models used for the simulation study in this chapter. The first joint model (Model 1) is the linear joint model having the exponential baseline hazard function. The second model (Model 2) is the penalized spline joint model having the Gompert distribution at baseline. In particular, Model 1 is of the form

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\} \\ &= \lambda \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}, \end{aligned} \tag{5.2.4}$$

where

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ m_i(t) &= \beta_0 + \beta_1 t + u_{i0} + u_{i1} t\,. \end{aligned} \tag{5.2.5}$$

Here, the random effects vector $\boldsymbol{b}_i = (u_{i0}, u_{i1})^T$ is assumed to have normal distribution with mean 0 and variance matrix $\boldsymbol{G} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$.

Model 2 has the form

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\} \\ &= \lambda_1 \exp(\lambda_2 t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}, \end{aligned} \tag{5.2.6}$$

where

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t) \\ m_i(t) &= \beta_0 + \beta_1 t + u_{i0} + u_{i1} t + \sum_{k=2}^{K} u_{ik} (t - \mathcal{K}_k)_+\,. \end{aligned} \tag{5.2.7}$$

Here, the random effects vector $\boldsymbol{b}_i = (u_{i0}, u_{i1}, u_{i2}, ..., u_{iK})^T$ is assumed to have a normal distribution with mean 0 and variance matrix $\boldsymbol{G} = Diag(G_{00}, G_{11}, G_{22}, ..., G_{KK})$. $\mathcal{K}_2$, $\mathcal{K}_3$,..., $\mathcal{K}_K$ are the fitted knots.

Following Wakefield (2013), the models can be rewritten in a hierarchical setting as follows:

Likelihood:

$$p(T_i, \delta_i, \boldsymbol{y}_i|\boldsymbol{\theta}, \boldsymbol{b}_i)\,, i = 1, .., n\,.$$

Random effects prior:

$$p(\boldsymbol{b}_i|\boldsymbol{G})\,, i = 1, .., n\,.$$

Hyperprior:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_t^T, \boldsymbol{\theta}_y^T, \boldsymbol{\theta}_b^T)^T\,.$$

5.3 Bayesian analysis

5.3.1 Prior distributions

Recall the full parameter vector of the joint model $\boldsymbol{\theta} = (\boldsymbol{\theta}_t^T, \boldsymbol{\theta}_y^T, \boldsymbol{\theta}_b^T)^T$. Following Wakefield (2013), Robert and Casella (2004) and Rizopoulos (2014), we assume the independence of the priors. The joint prior distribution, $p(\boldsymbol{\theta})$, can be written as

$$\begin{aligned} p(\boldsymbol{\theta}) &= p(\boldsymbol{\theta}_{h_0}, \boldsymbol{\gamma}, \alpha, \boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{G}) \\ &= p(\boldsymbol{\theta}_{h_0}) p(\boldsymbol{\gamma}, \alpha) p(\boldsymbol{\beta}, \sigma_\varepsilon^2) p(\boldsymbol{G}) \\ &= p(\boldsymbol{\theta}_{h_0}) p(\boldsymbol{\gamma}) p(\alpha) p(\boldsymbol{\beta}) p(\sigma_\varepsilon^2) p(\boldsymbol{G})\,. \end{aligned} \tag{5.3.1}$$

The individual prior distributions for survival part have the form

$$\begin{aligned} p(\boldsymbol{\theta}_{h_0}) &\sim \mathcal{MVN}(\mu_{\boldsymbol{\theta}_{h_0}}, \textstyle\sum_{\boldsymbol{\theta}_{h_0}})\,, \\ p(\boldsymbol{\gamma}) &\sim \mathcal{MVN}(\mu_\gamma, \textstyle\sum_\gamma)\,, \\ p(\alpha) &\sim \mathcal{UVN}(\mu_\alpha, \sigma_\alpha^2)\,. \end{aligned} \tag{5.3.2}$$

The individual prior distributions for longitudinal part have the form

$$\begin{aligned} p(\sigma_\varepsilon^2) &\sim \mathcal{IG}(a_0, b_0)\,, \\ p(\boldsymbol{\beta}) &\sim \mathcal{MVN}(\mu_\beta, \textstyle\sum_\beta)\,. \end{aligned} \tag{5.3.3}$$

Here, $\mathcal{IG}(a_0, b_0)$ is the inverse Gamma distribution with shape parameter a_0 and rate parameter b_0. In addition, the prior specification for $\boldsymbol{G}$ has two options according to Wakefield (2013). If $\boldsymbol{G}$ is a diagonal matrix with elements $\sigma_k^2, k = 1, ..., q$, the prior distribution for $\boldsymbol{G}$ has the form

$$p(\sigma_1^2, ..., \sigma_q^2) = \prod_{k=1}^{q} \mathcal{IG}(a_k, b_k)\,. \tag{5.3.4}$$

If $\boldsymbol{G}$ is a non-diagonal matrix, the inverse Wishart distribution is the conjugate prior for $\boldsymbol{G}$. In particular,

$$p(\boldsymbol{G}^{-1}) \sim \mathcal{IW}_{q+1}(r, \boldsymbol{R}^{-1})\,. \tag{5.3.5}$$

Here, $\mathcal{IW}_{q+1}(r, \boldsymbol{R}^{-1})$ is the inverse Wishart distribution with scale matrix $\boldsymbol{R}$ and r degrees of freedom. $(q+1)$ is the dimension of random effects.

5.3.2 Likelihood function

According to Section 3.3, the joint likelihood function for the penalized spline joint model has the form of

$$\begin{aligned}
p(T, \delta, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b}) &= \prod_{i=1}^{n} p(T_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{n_i} p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}) p(y_{ij}|\boldsymbol{b}_i; \boldsymbol{\theta}) \\
&= \prod_{i=1}^{n} \left[h_0(T_i) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}\right]^{\delta_i} \exp\left(-\int_0^{T_i} h_0(s) exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s) ds\right\}\right) \\
&\times \prod_{i=1}^{n} \prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{1}{2}}} \exp\left\{-\frac{(y_i(t_{ij}) - m_i(t_{ij}))^2}{2\sigma_\varepsilon^2}\right\}\,.
\end{aligned} \tag{5.3.6}$$

Here, we define the likelihood functions for Model 1 and Model 2 for later use in the simulation study.

For Model 1, the conditional density function for the survival part has the form

$$\begin{aligned}
p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) &= h(T_i \mid \mathcal{M}_i(T_i), \boldsymbol{w}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta})^{\delta_i} S(T_i \mid \mathcal{M}_i(T_i), \boldsymbol{w}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) \\
&= \left[h_0(T_i) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}\right]^{\delta_i} \\
&\times \exp\left(-\int_0^{T_i} h_0(s) exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s) ds\right\}\right),
\end{aligned}$$

where

$$m_i(t) = \beta_0 + \beta_1 t + u_{i0} + u_{i1} t \,.$$

The density function for the longitudinal part with the given random effects is

$$\begin{aligned} p(\boldsymbol{y_i}|\boldsymbol{b}_i;\boldsymbol{\theta}_y) &= \prod_j^{n_i} p\left\{y_i(t_{ij})|\boldsymbol{b_i};\boldsymbol{\theta}_y\right\} \\ &= \prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{1}{2}}} \exp\left\{-\frac{\left(y_i(t_{ij}) - (\beta_0 + \beta_1 t_{ij} + u_{i0} + u_{i1} t_{ij})\right)^2}{2\sigma_\varepsilon^2}\right\} . \end{aligned}$$

For Model 2, the conditional density function for the survival part has the form of

$$\begin{aligned} p(T_i, \delta_i \mid \boldsymbol{b}_i;\boldsymbol{\theta}_t,\boldsymbol{\beta}) &= h(T_i \mid \mathcal{M}_i(T_i), \boldsymbol{w}_i;\boldsymbol{\theta}_t,\boldsymbol{\beta})^{\delta_i} S(T_i \mid \mathcal{M}_i(T_i), \boldsymbol{w}_i;\boldsymbol{\theta}_t,\boldsymbol{\beta}) \\ &= \left[h_0(T_i)\exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_i + \alpha m_i(t)\right\}\right]^{\delta_i} \\ &\times \exp\left(-\int_0^{T_i} h_0(s) exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_i + \alpha m_i(s) ds\right\}\right), \end{aligned}$$

where

$$m_i(t) = \beta_0 + \beta_1 t + u_{i0} + u_{i1} t + \sum_{k=2}^{K} u_{ik}(t - \mathcal{K}_k)_+ \,.$$

The density function for the longitudinal part with the given random effects is

$$\begin{aligned} p(\boldsymbol{y_i}|\boldsymbol{b}_i;\boldsymbol{\theta}_y) &= \prod_j^{n_i} p\left\{y_i(t_{ij})|\boldsymbol{b_i};\boldsymbol{\theta}_y\right\} \\ &= \prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{1}{2}}} \exp\left\{-\frac{\left(y_i(t_{ij}) - \left(\beta_0 + \beta_1 t_{ij} + u_{i0} + u_{i1} t + \sum_{k=2}^{K} u_{ik}(t_{ij} - \mathcal{K}_k)_+\right)\right)^2}{2\sigma_\varepsilon^2}\right\} . \end{aligned}$$

5.3.3 Posterior distribution for the parameters

Using the prior distribution in (5.3.1) and the likelihood function in (5.3.6), the joint posterior distribution for the parameters $(\boldsymbol{\theta}, \boldsymbol{b})$ is obtained using

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{b}|T, \delta, \boldsymbol{y}) &\propto p(T, \delta, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b}) p(\boldsymbol{\theta}, \boldsymbol{b}) \\ &\propto \prod_{i=1}^{n} p(T_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{n_i} p(T_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{\theta}) p(y_{ij}\boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}_{h_0}) p(\boldsymbol{\gamma}) p(\alpha) p(\boldsymbol{\beta}) p(\sigma_\varepsilon^2) p(\boldsymbol{D}) \,. \end{aligned}$$

Applying (5.3.2), (5.3.3) and (5.3.4), the joint posterior distribution for $(\boldsymbol{\theta}, \boldsymbol{b})$ in Model 1 has the form

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{b}|T, \delta, \boldsymbol{y}) \propto & \prod_{i=1}^{n} \left[h_0(T_i) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}\right]^{\delta_i} \\
& \times \exp\left(-\int_0^{T_i} h_0(s) exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s) ds\right\}\right) \\
& \times \prod_{i=1}^{n}\prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{1}{2}}} \exp\left\{-\frac{\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i\right)^2}{2\sigma_\varepsilon^2}\right\} \\
& \times |\boldsymbol{G}|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{b}_i)^T \boldsymbol{G}(\boldsymbol{b}_i)\right\} \left|\sum_{\boldsymbol{\theta}_{h_0}}\right|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_{h_0} - \mu_{\boldsymbol{\theta}_{h_0}})^T \sum_{\boldsymbol{\theta}_{h_0}}^{-1} (\boldsymbol{\theta}_{h_0} - \mu_{\boldsymbol{\theta}_{h_0}})\right\} \\
& \times \left|\sum_{\gamma}\right|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma} - \mu_\gamma)^T \sum_{\lambda}^{-1} (\boldsymbol{\gamma} - \mu_\gamma)\right\} \\
& \times (2\pi\sigma_\alpha)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_\alpha^2}(\alpha - \mu_\alpha)^2\right\} \times \left|\sum_{\beta}\right|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mu_\beta)^T \sum_{\beta}^{-1} (\boldsymbol{\beta} - \mu_\beta)\right\} \\
& \times \frac{b_0^{a_0}}{\Gamma(a_0)} \left(\sigma_\varepsilon^2\right)^{-a_0 - 1} \exp(\frac{-b_0}{\sigma_\varepsilon^2}) \times |\boldsymbol{G}|^{(r-q)/2} \exp\left[-\frac{1}{2} tr(\boldsymbol{G}R)\right].
\end{aligned}
\tag{5.3.7}
$$

In a similar way, using (5.3.2), (5.3.3) and (5.3.5), the joint posterior distribution for $(\boldsymbol{\theta}, \boldsymbol{b})$ in Model 2 has the form

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{b}|T, \delta, \boldsymbol{y}) \propto & \prod_{i=1}^{n} \left[h_0(T_i) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}\right]^{\delta_i} \\
& \times \prod_{i=1}^{n} \exp\left(-\int_0^{T_i} h_0(s) exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s) ds\right\}\right) \\
& \times \prod_{i=1}^{n}\prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{1}{2}}} \exp\left\{-\frac{\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i\right)^2}{2\sigma_\varepsilon^2}\right\} \\
& \times |\boldsymbol{G}|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{b}_i)^T \boldsymbol{G}(\boldsymbol{b}_i)\right\} \left|\sum_{\boldsymbol{\theta}_{h_0}}\right|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_{h_0} - \mu_{\boldsymbol{\theta}_{h_0}})^T \sum_{\boldsymbol{\theta}_{h_0}}^{-1} (\boldsymbol{\theta}_{h_0} - \mu_{\boldsymbol{\theta}_{h_0}})\right\} \\
& \times \left|\sum_{\gamma}\right|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma} - \mu_\gamma)^T \sum_{\lambda}^{-1} (\boldsymbol{\gamma} - \mu_\gamma)\right\} \\
& \times (2\pi\sigma_\alpha)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_\alpha^2}(\alpha - \mu_\alpha)^2\right\} \left|\sum_{\beta}\right|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mu_\beta)^T \sum_{\beta}^{-1} (\boldsymbol{\beta} - \mu_\beta)\right\} \\
& \times \frac{b_0^{a_0}}{\Gamma(a_0)} \left(\sigma_\varepsilon^2\right)^{-a_0 - 1} \exp(\frac{-b_0}{\sigma_\varepsilon^2}) \times \prod_{k=1}^{q} {\tau_k}^{a_k - 1} \exp(-b_k \tau_k),
\end{aligned}
\tag{5.3.8}
$$

where $\tau_k = 1/\sigma_k^2$ for $k = 1, ..., q$.

The joint posterior distributions for $(\boldsymbol{\theta}, \boldsymbol{b})$ in (5.3.7) and (5.3.8) are not in the standard forms. Therefore, it is difficult to sample from the joint posterior distribution directly.

In order to sample each parameter from the joint model, we need to define the target distribution for each of the parameters using their conditional posterior distributions. Based on the joint posterior distribution in (5.3.7) and (5.3.8), the conditional posterior distribution for each of the parameters in the joint models is derived as follows.

Using (5.3.7), the conditional posterior distribution for the parameters in the baseline hazard function, $\boldsymbol{\theta}_{h_0}$, has the form

$$\begin{aligned} p(\boldsymbol{\theta}_{h_0}|\boldsymbol{\theta}_{\left(-\boldsymbol{\theta}_{h_0}\right)}, \boldsymbol{b}, T, \delta, \boldsymbol{y}) \propto & \prod_{i=1}^{n} \left[h_0(T_i) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}\right]^{\delta_i} \\ & \times \prod_{i=1}^{n} \exp\left(-\int_0^{T_i} h_0(s) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s) ds\right\}\right) \\ & \left|\sum_{\boldsymbol{\theta}_{h_0}}\right|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_{h_0} - \mu_{\boldsymbol{\theta}_{h_0}})^T \sum_{\boldsymbol{\theta}_{h_0}}^{-1} (\boldsymbol{\theta}_{h_0} - \mu_{\boldsymbol{\theta}_{h_0}})\right\}, \end{aligned} \tag{5.3.9}$$

where the notation $\boldsymbol{\theta}_{(-\theta_i)}$ means all of the parameters in the joint model except for θ_i.

Using (5.3.7), the conditional posterior distribution for the regression coefficients, $(\boldsymbol{\gamma}, \alpha)$, in the survival submodel has the form

$$\begin{aligned} p(\boldsymbol{\gamma}, \alpha|\boldsymbol{\theta}_{(-\boldsymbol{\gamma},-\alpha)}, \boldsymbol{b}, T, \delta, \boldsymbol{y}) \propto & \prod_{i=1}^{n} \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}^{\delta_i} \exp\left(-\int_0^{T_i} exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s) ds\right\}\right) \\ & \times \left|\sum_{\boldsymbol{\gamma}}\right|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma} - \mu_\gamma)^T \sum_{\lambda}^{-1} (\boldsymbol{\gamma} - \mu_\gamma)\right\} \\ & \times (2\pi\sigma_\alpha)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_\alpha^2}(\alpha - \mu_\alpha)^2\right\}. \end{aligned} \tag{5.3.10}$$

In a similar way, the conditional posterior distribution for the regression coefficients $\boldsymbol{\beta}$ in the linear mixed effects submodel has the form

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{\theta}_{(-\beta)}, \boldsymbol{b}, T, \delta, \boldsymbol{y}) \propto & \prod_{i=1}^{n} \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}^{\delta_i} \exp\left(-\int_0^{T_i} exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(s) ds\right\}\right) \\ & \times \prod_{i=1}^{n} \prod_{j=1}^{n_i} \exp\left\{-\frac{\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i\right)^2}{2\sigma_\varepsilon^2}\right\} \\ & \times \left|\sum_{\beta}\right|^{\frac{-1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mu_\beta)^T \sum_{\beta}^{-1} (\boldsymbol{\beta} - \mu_\beta)\right\}. \end{aligned} \tag{5.3.11}$$

The prior distribution of σ_ε^2 is an inverse gamma distribution with a scale of a_0 and a shape of b_0 as in (5.3.3), which is the conjugate prior distribution. In particular, by setting $\tau = 1/\sigma_\varepsilon^2$ and $N = \sum_{i=1}^{n} n_i$, the posterior distribution of τ is proportional to

$$p(\tau|\boldsymbol{\theta}_{(-\tau)}, \boldsymbol{b}, T, \delta, \boldsymbol{y}) \propto \prod_{i=1}^{n}\prod_{j=1}^{n_i}\frac{1}{\sigma_\varepsilon}\exp\left\{-\frac{\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i\right)^2}{2\sigma_\varepsilon^2}\right\}$$
$$\times \tau^{a_0-1}\exp(-b_0\tau)$$
$$\propto \tau^{N/2+a_0-1}\exp(-b_0\tau)$$
$$\times \exp\left\{-\frac{\tau\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n_i}\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i\right)^2}{2}\right\}. \tag{5.3.12}$$

The result is that the conditional posterior distribution of $\tau = 1/\sigma_\varepsilon^2$ is distributed as $\mathcal{G}(\alpha^*, \beta^*)$ where

$$\alpha^* = a_0 + \frac{N}{2},$$
$$\beta^* = b_0 + \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n_i}\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i\right)^2}{2}.$$

The conditional posterior distribution for random effects $\boldsymbol{b}_i$ in the linear mixed effects submodel has the form

$$p(\boldsymbol{b}_i|\boldsymbol{\theta}_{(-\boldsymbol{b}_i)}, T, \delta, \boldsymbol{y}) \propto \exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_i + \alpha m_i(t)\right\}^{\delta_i}\exp\left(-\int_0^{T_i} exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_i + \alpha m_i(s)ds\right\}\right)$$
$$\times \prod_{j=1}^{n_i}\frac{1}{(2\pi\sigma_\varepsilon^2)^{\frac{1}{2}}}\exp\left\{-\frac{\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \left[\begin{array}{cc}\boldsymbol{X}_i^T(t_{ij}) & \boldsymbol{Z}_i^T(t_{ij})\end{array}\right]\boldsymbol{b}_i\right)^2}{2\sigma_\varepsilon^2}\right\}$$
$$\times |\boldsymbol{G}|^{\frac{-1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{b}_i)^T\boldsymbol{G}(\boldsymbol{b}_i)\right\}. \tag{5.3.13}$$

There are two options for choosing the conjugate prior distribution of variance matrix $\boldsymbol{G}$. These are when $\boldsymbol{G}$ is a diagonal matrix and when $\boldsymbol{G}$ is a non-diagonal matrix. Therefore, we propose two conditional posterior distributions for matrix $\boldsymbol{G}$.

In the case when $\boldsymbol{G}$ is a non-diagonal matrix, the conjugate prior distribution for $\boldsymbol{G}$ is a Wishart distribution as in (5.3.5). In particular, the distribution of $\boldsymbol{G}^{-1}$ has the form

$$p(\boldsymbol{G}^{-1}) = c^{-1}|\boldsymbol{G}^{-1}|^{(r-q)/2}\exp\left[-\frac{1}{2}tr(\boldsymbol{G}^{-1}\boldsymbol{R})\right], \tag{5.3.14}$$

where

$$c = 2^{r(q+1)/2}\Gamma_q(r/2)|R^{-1}|^{r/2},$$

with

$$\Gamma_q(r/2) = \pi^{(q-1)q/4} \prod_{j=1}^{q} \Gamma[(r+1-j)/2]\,.$$

Using (5.3.14), the conditional posterior distribution for $\boldsymbol{G}^{-1}$ is proportional to

$$\begin{aligned} p(\boldsymbol{G}^{-1}|\boldsymbol{\theta}_{\left(-\boldsymbol{G}^{-1}\right)}, \boldsymbol{b}, T, \delta, \boldsymbol{y}) &\propto \prod_{i=1}^{n} p(b_i|\boldsymbol{G}^{-1}) p(\boldsymbol{G}^{-1}) \\ &\propto \prod_{i=1}^{n} |\boldsymbol{G}|^{\frac{-1}{2}} \exp(-\frac{1}{2} b_i^T \boldsymbol{G}^{-1} b_i) |\boldsymbol{G}^{-1}|^{(r-q)/2} \exp\left[-\frac{1}{2} tr(\boldsymbol{G}^{-1}\boldsymbol{R})\right] \\ &= |\boldsymbol{G}^{-1}|^{(n+r-q)/2} \exp\left\{-\frac{1}{2} tr\left(\boldsymbol{G}^{-1}\left[\sum_{i=1}^{n}(b_i b_i^T) + \boldsymbol{R}\right]\right)\right\}. \end{aligned} \tag{5.3.15}$$

Moreover,

$$b_i^T \boldsymbol{G}^{-1} b_i = tr[b_i^T \boldsymbol{G}^{-1} b_i] = tr[\boldsymbol{G}^{-1} b_i b_i^T]. \tag{5.3.16}$$

From (5.3.15) and (5.3.16), we have

$$p(\boldsymbol{G}^{-1}|\boldsymbol{\theta}_{\left(-\boldsymbol{G}^{-1}\right)}, \boldsymbol{b}, T, \delta, \boldsymbol{y}) \propto |\boldsymbol{G}^{-1}|^{(n+r-q)/2} \exp\left\{-\frac{1}{2} tr\left(\boldsymbol{G}^{-1}\left[\sum_{i=1}^{n}(b_i b_i^T) + \boldsymbol{R}\right]\right)\right\}. \tag{5.3.17}$$

The conditional posterior distribution for $\boldsymbol{G}^{-1}$ has the standard form

$$\boldsymbol{G}^{-1}|\boldsymbol{\theta}_{\left(-\boldsymbol{G}^{-1}\right)}, \boldsymbol{b}, T, \delta, \boldsymbol{y} \sim \mathcal{W}_q\left[n+r, \left(\boldsymbol{R} + \sum_{i=1}^{n}(b_i b_i^T)\right)^{-1}\right]. \tag{5.3.18}$$

In the second case when $\boldsymbol{G}$ is a diagonal matrix with elements σ_k^2, $k = 1, ..., q$, the conjugate prior distribution for $\boldsymbol{G}$ has the form

$$p(\sigma_1^2, ..., \sigma_q^2) = \prod_{k=1}^{q} \mathcal{IG}(a_k, b_k)\,.$$

Set $\tau_k = 1/\sigma_k^2$ for $k = 1, ..., q$. The conditional posterior distribution for $\boldsymbol{G}^{-1}$ has the standard form

$$\begin{aligned} p(\boldsymbol{G}^{-1}|\boldsymbol{\theta}_{\left(-\boldsymbol{G}^{-1}\right)}, \boldsymbol{b}, T, \delta, \boldsymbol{y}) &= p\left(\tau_1, ..., \tau_q|\boldsymbol{\theta}_{(-\tau_1,...,-\tau_q)}, T, \delta, \boldsymbol{y}\right) \\ &\propto \prod_{i=1}^{n} p(b_i|\boldsymbol{G}^{-1}) \prod_{k=1}^{q} p(\tau_k) \\ &\propto \prod_{i=1}^{n} |\boldsymbol{G}|^{\frac{-1}{2}} \exp(-\frac{1}{2} b_i^T \boldsymbol{D}^{-1} b_i) \prod_{k=1}^{q} {\tau_k}^{a_k - 1} \exp(-b_k \tau_k) \\ &= \prod_{k=1}^{q} {\tau_k}^{n/2 + a_k - 1} \exp(-b_k \tau_k) \exp\left\{\sum_{i=1}^{n}(-\frac{1}{2} b_i^T b_i \tau_k)\right\} \\ &= \prod_{k=1}^{q} {\tau_k}^{n/2 + a_k - 1} \exp\left\{-\left(b_k + \frac{1}{2}\sum_{i=1}^{n} b_i^T b_i\right)\tau_k\right\}. \end{aligned} \tag{5.3.19}$$

Therefore,

$$\tau_1, ..., \tau_q | \boldsymbol{\theta}_{(-\tau_1,...,-\tau_q)}, T, \delta, \boldsymbol{y} \sim \prod_{k=1}^{q} \mathcal{G}(a_k^*, b_k^*) \,,$$

where

$$a_k^* = a_k + \frac{n}{2} \,,$$
$$b_k^* = b_k + \frac{1}{2} \sum_{i=1}^{n} b_i^T b_i \,.$$

In summary, we have four groups of parameter vectors having non-standard conditional posterior distributions. They are the conditional posterior distributions for parameters in the baseline hazard function, $\boldsymbol{\theta}_{h_0}$, the regression coefficients in the survival submodel, $(\boldsymbol{\gamma}, \alpha)$, the regression coefficients in the linear mixed effects submodel, $\boldsymbol{\beta}$, and the random effects in the linear mixed effects submodel, $\boldsymbol{b}_i$. We also have two groups of parameter vectors having standard conditional posterior distributions. They are the conditional posterior distributions for the parameters of error terms, σ_ε^2, and variance matrix of random effects, $\boldsymbol{G}$. Henceforth, to sample for parameters in the joint model in (5.2.1), we implement the MH algorithm for the four groups of parameters vectors having non-standard conditional posterior distributions and the GS algorithm for the two groups of the parameter vectors having standard conditional posterior distributions.

5.4 The main algorithm

In order to simulate the parameters of the joint model, a set of MCMC algorithms is implemented in the proposed main algorithm. In particular, the GS algorithms are employed using the standard conditional posterior distributions from Section 5.3.3. For parameters having a non-standard conditional posterior distribution, we use MH algorithms. The main algorithm for the penalized spline joint model is described. Some sub-algorithms will be detailed separately.

From the joint models in (5.2.4) and (5.2.6), the hazard rate at baseline directly and significantly affects the hazard rate of each subject. Moreover, the baseline risk function $h_0(.)$ is unspecified. In order to choose noninformative priors for the parameters of the hazard rate at baseline, independent gamma distributions are implemented (Ibrahim et al., 2005; Brown et al., 2005). However, for these chosen prior distributions, 400,000 iterations were required to achieve convergence with a burning-in of 200,000 iterations. To reduce

Main Algorithm

(1) Initialise $\boldsymbol{\theta}_t^{(0)} = (\boldsymbol{\theta}_{h_0}^{(0)}, \boldsymbol{\gamma}^{(0)}, \alpha^{(0)})^T$, $\boldsymbol{\theta}_y^{(0)} = (\boldsymbol{\beta}^{(0)}, \sigma_\varepsilon^2)^T$,$\boldsymbol{\theta}_b^{(0)} = \boldsymbol{G}^{(0)}$ and $\boldsymbol{b}^{(0)}$ either randomly or deterministically

(2) for $t = 1$ to T do

Given the current parameters of $\boldsymbol{\theta}_t^{(t-1)}$,$\boldsymbol{\theta}_y^{(t-1)}$,$\boldsymbol{\theta}_b^{(t-1)}$ and $\boldsymbol{b}^{(t-1)}$

(2.1) Simulate the parameters for survival part:

(i) $\boldsymbol{MH}_{\boldsymbol{\theta}_{h_0}}$ step. Simulate the parameter vector, $\boldsymbol{\theta}_{h_0}$, in the baseline hazard function using (5.3.9):

$$(\boldsymbol{\theta}_{h_0}^{(t)}) \sim p(\boldsymbol{\theta}_{h_0}|\boldsymbol{\gamma}^{(t-1)}, \alpha^{(t-1)}, \boldsymbol{\beta}^{(t-i)}, \sigma_\varepsilon^{2(t-1)}, \boldsymbol{G}^{(t-1)}, \boldsymbol{b}^{(t-1)}, T, \delta, \boldsymbol{y})$$

(ii)$\boldsymbol{MH}_{\boldsymbol{\gamma},\alpha}$ step. Simulate the parameter vector of the regression coefficients, $(\boldsymbol{\gamma}, \alpha)$, in the survival submodel using (5.3.10):

$$(\boldsymbol{\gamma}, \alpha)^{(t)} \sim p(\boldsymbol{\gamma}, \alpha|\boldsymbol{\theta}_{h_0}^{(t)}, \boldsymbol{\beta}^{(t-i)}, \sigma_\varepsilon^{2(t-1)}, \boldsymbol{G}^{(t-1)}, \boldsymbol{b}^{(t-1)}, T, \delta, \boldsymbol{y})$$

(2.2) Simulate the parameters for the longitudinal part:

(i) $\boldsymbol{MH}_{\boldsymbol{\beta}}$ step. Simulate the parameter vector of the regression coefficients, $\boldsymbol{\beta}$, in the longitudinal submodel using (5.3.11):

$$(\boldsymbol{\beta}^{(t)}) \sim p(\boldsymbol{\beta}|\boldsymbol{\theta}_{h_0}^{(t)}, \boldsymbol{\gamma}^{(t)}, \alpha^{(t)}, \sigma_\varepsilon^{2(t-1)}, \boldsymbol{G}^{(t-1)}, \boldsymbol{b}^{(t-1)}, T, \delta, \boldsymbol{y})$$

(ii)$\boldsymbol{GS}_{\sigma_\varepsilon^2}$ step. Simulate the variance of error measurement, σ_ε^2, using (5.3.12):

$$(\sigma_\varepsilon^{2(t)}) \sim p(\sigma_\varepsilon^2|\boldsymbol{\theta}_{h_0}^{(t)}, \boldsymbol{\gamma}^{(t)}, \alpha^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{G}^{(t-1)}, \boldsymbol{b}^{(t-1)}, T, \delta, \boldsymbol{y})$$

(2.3) Simulate the parameters for the random effects as follows

(i)$\boldsymbol{MH}_{\boldsymbol{b}}$ step. Simulate the random effects using (5.3.13):

$$(\boldsymbol{b}^{(t)}) \sim p(\boldsymbol{b}|\boldsymbol{\theta}_{h_0}^{(t)}, \boldsymbol{\gamma}^{(t)}, \alpha^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{G}^{(t-1)}, \sigma_\varepsilon^{2(t)}, T, \delta, \boldsymbol{y})$$

(ii) $\boldsymbol{GS}_{\boldsymbol{G}}$ step. Simulate the variance matrix of the random effects using (5.3.18) or (5.3.19):

$$(\boldsymbol{G}^{(t)}) \sim p(\boldsymbol{G}|\boldsymbol{\theta}_{h_0}^{(t)}, \boldsymbol{\gamma}^{(t)}, \alpha^{(t)}, \boldsymbol{\beta}^{(t)}, \sigma_\varepsilon^{2(t)}, \boldsymbol{b}^{(t)}, T, \delta, \boldsymbol{y})$$

(3) end for

the size of MCMC sampling and achieve a faster convergence, we take advantage of the fitted model for longitudinal data using the linear mixed effects model and the fitted model for survival data using the Cox model.

First, in R, we run the lme and coxph functions separately for longitudinal data and survival data. The estimated values from these functions are used to define the parameters of the empirical Bayes prior distributions. In particular, the prior distributions for the parameters in $h_0(t)$ have normal distributions with mean and variance defined from coxph function. In addition to this, from these estimated values from the linear mixed effects model and survival model, we define the means for the normal prior distributions of the regression coefficients of the survival submodel and longitudinal submodel. To ensure the flexibility as flat priors, the variances for the normal prior distributions are set 100 times

the estimated variances using the lme and coxph functions.

To support the main algorithm, the MH steps for the parameters of the baseline hazard function, the regression coefficients in the survival submodel, the regression coefficients in the longitudinal submodel and random effects will be presented in detail in the following sections. In addition, the acceptance probabilities and the proposal distribution will also be addressed in these sub-algorithms.

5.4.1 $MH_{\boldsymbol{\theta}_{h_0}}$ step

Because the hazard baseline is considered to be unspecified in the joint model, we apply the two-stage approach to identify the prior distribution for the parameters. In particular, we first run a single MH algorithm to simulate samples for $\boldsymbol{\theta}_{h_0}$ using the estimated values from the linear mixed effects model and Cox model to choose a weakly informative priors for the parameters. The mean of the priors for the baseline hazard parameters are the mean of the MCMC samples. The variance of the priors for the baseline hazard parameters is set at 100 times the variance of the MCMC samples. The MH acceptance ratio for $\boldsymbol{\theta}_{h_0}$ is then calculated as

$$r_{\boldsymbol{\theta}_{h_0}} = \frac{p(\boldsymbol{\theta}_{h_0}^{(prop)}|\boldsymbol{\theta}_{(-\boldsymbol{\theta}_{h_0})}, \boldsymbol{b}, T, \delta, y) q(\boldsymbol{\theta}_{h_0}^{(curr)}|\boldsymbol{\theta}_{h_0}^{(prop)})}{p(\boldsymbol{\theta}_{h_0}^{(curr)}|\boldsymbol{\theta}_{(-\boldsymbol{\theta}_{h_0})}, \boldsymbol{b}, T, \delta, y) q(\boldsymbol{\theta}_{h_0}^{(prop)}|\boldsymbol{\theta}_{h_0}^{(curr)})}\,. \tag{5.4.1}$$

Here, $\boldsymbol{\theta}_{h_0}^{(prop)}$and $\boldsymbol{\theta}_{h_0}^{(curr)}$ are the proposed and current values of $\boldsymbol{\theta}_{h_0}$ respectively and $p(\boldsymbol{\theta}_{h_0}|\boldsymbol{\theta}_{(-\boldsymbol{\theta}_{h_0})}, \boldsymbol{b}, T, \delta, y)$ is the conditional posterior distribution for $\boldsymbol{\theta}_{h_0}$ as in (5.3.9). Moreover, $q(\boldsymbol{\theta}_{h_0}^{(prop)}|\boldsymbol{\theta}_{h_0}^{(curr)})$ is the proposal density for the baseline hazard parameter vector, $\boldsymbol{\theta}_{h_0}$. We employ a multivariate normal distribution centred at the current value of $\boldsymbol{\theta}_{h_0}$as the proposal density for the baseline hazard parameter vector. The proposal distribution for $\boldsymbol{\theta}_{h_0}$ has the form

$$q(\boldsymbol{\theta}_{h_0}^{(prop)}|\boldsymbol{\theta}_{h_0}^{(curr)}) \sim \mathcal{MVN}(\boldsymbol{\theta}_{h_0}^{(curr)}, \triangle_{\boldsymbol{\theta}_{h_0}} \boldsymbol{I})\,, \tag{5.4.2}$$

where $\triangle_{\boldsymbol{\theta}_{h_0}}$ is a tuning parameter and $\boldsymbol{I}$ is an identity matrix. With this choice for the proposal distribution, the algorithm becomes the random walk MH algorithm (Geman and Geman, 1984; Gelman and Hill, 2007). As a result, the ratio for the proposal densities is always one.

In the MH algorithm, the acceptance rate is one of the factors to check for the convergence of a chain (Gelman et al., 1995). If the acceptance rate is too high, this means that most of the proposed values are accepted. This can lead to a wiggle trace plot for a chain and can be time consuming sampling for the entire parameter space. In the other hand, if the acceptance rate is too low, only a few of the proposed values are accepted. The sample will stay at the same level or it has large jumps. This can affect the convergence of a chain. Therefore, a tuning parameter, $\triangle_{\boldsymbol{\theta}_{h_0}}$, is chosen so that the desirable acceptance rate is between 20% and 50 %

Wakefield (2013); Robert and Casella (2004).

The MH acceptance ratio has the form

$$\begin{aligned} r_{\boldsymbol{\theta}_{h_0}} = & \frac{\prod_{i=1}^{n}\left[\left\{h_0(T_i)|\boldsymbol{\theta}_{h_0}^{(prop)}\right\}\exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_i+\alpha m_i(T_i)\right\}\right]^{\delta_i}}{\prod_{i=1}^{n}\left[\left\{h_0(T_i)|\boldsymbol{\theta}_{h_0}^{(curr)}\right\}\exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_i+\alpha m_i(T_i)\right\}\right]^{\delta_i}} \\ & \times\prod_{i=1}^{n}\exp\left(-\int_0^{T_i}\left\{h_0(s)|\boldsymbol{\theta}_{h_0}^{(prop)}-h_0(s)|\boldsymbol{\theta}_{h_0}^{(curr)}\right\}\exp\left\{\boldsymbol{\gamma}^T\boldsymbol{w}_i+\alpha m_i(s)\right\}ds\right) \\ & \times\frac{\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_{h_0}^{(prop)}-\mu_{\boldsymbol{\theta}_{h_0}})^T\sum_{\boldsymbol{\theta}_{h_0}}^{-1}(\boldsymbol{\theta}_{h_0}^{(prop)}-\mu_{\boldsymbol{\theta}_{h_0}})\right\}}{\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_{h_0}^{(curr)}-\mu_{\boldsymbol{\theta}_{h_0}})^T\sum_{\boldsymbol{\theta}_{h_0}}^{-1}(\boldsymbol{\theta}_{h_0}^{(curr)}-\mu_{\boldsymbol{\theta}_{h_0}})\right\}}. \end{aligned} \tag{5.4.3}$$

The sub-algorithm for the parameters at the baseline hazard $\boldsymbol{\theta}_{h_0}$ is outlined as follows:

$\boldsymbol{MH}_{\boldsymbol{\theta}_{h_0}}$ step: The single MH for hazard rate at baseline

1. Given the current state $\left(\boldsymbol{\theta}_{h_0}^{(t-1)},\boldsymbol{\gamma}^{(t-1)},\alpha^{(t-1)},\boldsymbol{\beta}^{(t-i)},\sigma_{\varepsilon}^{2(t-1)},\boldsymbol{G}^{(t-1)},\boldsymbol{b}^{(t-1)}\right)$
2. Propose a new parameter vector,$\boldsymbol{\theta}_{h_0}^{(prop)}$, from proposal distribution as in (5.4.2)
3. Calculate the acceptance probability, $r_{\boldsymbol{\theta}_{h_0}}$, using (5.4.3)
4. Simulate $u\sim\mathcal{U}(0,1)$:
   i. If $r_{\boldsymbol{\theta}_{h_0}}>u$, then set $\boldsymbol{\theta}_{h_0}^{(t)}=\boldsymbol{\theta}_{h_0}^{(prop)}$
   ii. Else, set $\boldsymbol{\theta}_{h_0}^{(t)}=\boldsymbol{\theta}_{h_0}^{(t-1)}$

5.4.2 $\boldsymbol{MH}_{(\boldsymbol{\gamma},\alpha)}$ step

The MH acceptance ratio for the parameters $(\boldsymbol{\gamma},\alpha)$ has the form

$$r_{(\gamma,\alpha)}=\frac{p(\boldsymbol{\gamma}^{(prop)},\alpha^{(prop)}|\boldsymbol{\theta}_{(-\gamma,-\alpha)},\boldsymbol{b},T,\delta,y)q(\boldsymbol{\gamma}^{(curr)}|\boldsymbol{\gamma}^{(prop)})q(\alpha^{(curr)}|\alpha^{(prop)})}{p(\boldsymbol{\gamma}^{(curr)},\alpha^{(curr)}|\boldsymbol{\theta}_{(-\gamma,-\alpha)},\boldsymbol{b},T,\delta,y)q(\boldsymbol{\gamma}^{(prop)}|\boldsymbol{\gamma}^{(curr)})q(\alpha^{(prop)}|\alpha^{(curr)})}. \tag{5.4.4}$$

Here, $\boldsymbol{\gamma}^{(prop)}$, $\alpha^{(prop)}$, $\boldsymbol{\gamma}^{(curr)}$ and $\alpha^{(curr)}$ are the proposed and current values of the parameter vector $\boldsymbol{\gamma}$ and α respectively. The notation $p(\boldsymbol{\gamma}, \alpha|\boldsymbol{\theta}_{(-\boldsymbol{\gamma},-\boldsymbol{\alpha})}, \boldsymbol{b})$ is the conditional posterior distribution for $\boldsymbol{\gamma}$ and α as in (5.3.10). Moreover, $q(\boldsymbol{\gamma}^{(prop)}|\boldsymbol{\gamma}^{(curr)})$ and $q(\alpha^{(prop)}|\alpha^{(curr)})$ are the proposal densities for the coefficient vector of the survival part. The proposal distributions for the parameter vector are chosen as follows

$$\begin{aligned} q(\boldsymbol{\gamma}^{(prop)}|\boldsymbol{\gamma}^{(curr)}) &\sim \mathcal{MVN}(\boldsymbol{\gamma}^{(curr)}, \triangle_{\gamma}\hat{\boldsymbol{V}})\,, \\ q(\alpha^{(prop)}|\alpha^{(curr)}) &\sim \mathcal{UVN}(\alpha^{(curr)}, \triangle_{\alpha})\,, \end{aligned} \tag{5.4.5}$$

where $\hat{\boldsymbol{V}}$ is the asymptotic variance-covariance matrix of the parameter vector $\boldsymbol{\gamma}$ from the Cox model, $\triangle_{\gamma}$ and $\triangle_{\alpha}$ are the tuning parameters. Based on the conditional posterior in (5.3.10) and the proposal distribution in (5.4.5), the MH acceptance ratio for $\boldsymbol{\gamma}$ and α has the form

$$\begin{aligned} r_{(\boldsymbol{\gamma},\alpha)} = & \frac{\prod_{i=1}^{n}\left[\exp\left\{(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(prop)})\boldsymbol{w}_i + \alpha^{(prop)}m_i(T_i)\right\}\right]^{\delta_i}}{\prod_{i=1}^{n}\left[\exp\left\{(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(curr)})\boldsymbol{w}_i + \alpha^{(curr)}m_i(T_i)\right\}\right]^{\delta_i}} \\ & \times \frac{\prod_{i=1}^{n}\exp\left(-\int_0^{T_i} exp\left\{(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(prop)})\boldsymbol{w}_i + \alpha^{(prop)}m_i(s)ds\right\}\right)}{\prod_{i=1}^{n}\exp\left(-\int_0^{T_i} exp\left\{(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(curr)})\boldsymbol{w}_i + \alpha^{(curr)}m_i(s)ds\right\}\right)} \\ & \times \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{\gamma}^{(prop)} - \mu_{\gamma})^T\sum_{\lambda}^{-1}(\boldsymbol{\gamma}^{(prop)} - \mu_{\gamma})\right\}}{\exp\left\{-\frac{1}{2}(\boldsymbol{\gamma}^{(curr)} - \mu_{\gamma})^T\sum_{\lambda}^{-1}(\boldsymbol{\gamma}^{(curr)} - \mu_{\gamma})\right\}} \\ & \times \exp\left\{-\frac{(\alpha^{prop} - \mu_{\alpha})^2}{2\sigma_{\alpha}^2} + \frac{(\alpha^{curr} - \mu_{\alpha})^2}{2\sigma_{\alpha}^2}\right\}. \end{aligned} \tag{5.4.6}$$

The sub-algorithm for the parameters $(\boldsymbol{\gamma}, \alpha)$ is now detailed as follows.

$\boldsymbol{MH}_{(\gamma,\alpha)}$: The single MH samplers for survival coefficients

1. Given the current state $\left(\boldsymbol{\theta}_{h_0}^{(t)}, \boldsymbol{\gamma}^{(t-1)}, \alpha^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \sigma_{\varepsilon}^{2(t-1)}, \boldsymbol{G}^{(t-1)}, \boldsymbol{b}^{(t-1)}\right)$
2. Propose new values for the vector $\boldsymbol{\gamma}^{(prop)}$ from proposal distribution as in (5.4.5)
3. Propose new values for the vector $\alpha^{(prop)}$ from proposal distribution as in (5.4.5)
4. Calculate the acceptance probability, $r_{(\gamma,\alpha)}$, using (5.4.6)
5. Simulate $u \sim \mathcal{U}(0,1)$:
   i. If $r_{(\gamma,\alpha)} > u$, then set $(\boldsymbol{\gamma}^{(t)}, \alpha^{(t)}) = (\boldsymbol{\gamma}^{(prop)}, \alpha^{(prop)})$
   ii. Else, set $(\boldsymbol{\gamma}^{(t)}, \alpha^{(t)}) = (\boldsymbol{\gamma}^{(t-1)}, \alpha^{(t-1)})$

5.4.3 MH_β step

In a similar way, the MH acceptance ratio for the parameter vector $\boldsymbol{\beta}$ has the form

$$r_\beta = \frac{p(\boldsymbol{\beta}^{(prop)}|\theta_{(-\beta)}, \boldsymbol{b}, T, \delta, y) q(\boldsymbol{\beta}^{(curr)}|\boldsymbol{\beta}^{(prop)})}{p(\boldsymbol{\beta}^{(curr)}|\theta_{-\beta_i}, \boldsymbol{b}, T, \delta, y) q(\boldsymbol{\beta}^{(prop)}|\boldsymbol{\beta}^{(curr)})} \,. \tag{5.4.7}$$

Here, $\boldsymbol{\beta}^{(prop)}$ and $\boldsymbol{\beta}^{(curr)}$ are the proposed and current values of the parameter vector $\boldsymbol{\beta}$ respectively. The notation $p(\boldsymbol{\beta}|\boldsymbol{\theta}_{(-\beta)}, \boldsymbol{b}, T, \delta, y)$ is the conditional posterior distribution for $\boldsymbol{\beta}$ as in (5.3.11). The proposal density for the coefficient vector of the longitudinal part, $q(\boldsymbol{\beta}^{(prop)}|\boldsymbol{\beta}^{(curr)})$ is chosen as a multivariate normal distribution

$$q(\boldsymbol{\beta}^{(prop)}|\boldsymbol{\beta}^{(curr)}) \sim \mathcal{MVN}(\boldsymbol{\beta}^{(curr)}, \triangle_\beta \hat{\boldsymbol{W}}) \,, \tag{5.4.8}$$

where $\hat{\boldsymbol{W}}$ is the asymptotic variance-covariance matrix of the parameter vector $\boldsymbol{\gamma}$ from the linear mixed effects model and $\triangle_\beta$ is the tuning parameter. Based on the conditional posterior in (5.3.11) and the proposal distribution in (5.4.8), the MH acceptance ratio for the parameter vector $\boldsymbol{\beta}$ has the form

$$\begin{aligned}
r_\beta = & \frac{\prod_{i=1}^{n} \left[\exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha\left(m_i(T_i)|\boldsymbol{\beta}^{(prop)}\right)\right\}\right]^{\delta_i}}{\prod_{i=1}^{n} \left[\exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha\left(m_i(T_i)|\boldsymbol{\beta}^{(curr)}\right)\right\}\right]^{\delta_i}} \\
& \times \frac{\prod_{i=1}^{n} \exp\left(-\int_0^{T_i} exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha\left(m_i(s)|\boldsymbol{\beta}^{(prop)}\right) ds\right\}\right)}{\prod_{i=1}^{n} \exp\left(-\int_0^{T_i} exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha\left(m_i(s)|\boldsymbol{\beta}^{(curr)}\right) ds\right\}\right)} \\
& \times \frac{\prod_{i=1}^{n}\prod_{j=1}^{n_i} \exp\left\{-\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta}^{(prop)} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i\right)^2 / 2\sigma_\varepsilon^2\right\}}{\prod_{i=1}^{n}\prod_{j=1}^{n_i} \exp\left\{-\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta}^{(curr)} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i\right)^2 / 2\sigma_\varepsilon^2\right\}} \\
& \times \frac{\exp\left\{-(\boldsymbol{\beta}^{(prop)} - \mu_\beta)^T \sum_\beta^{-1} (\boldsymbol{\beta}^{(prop)} - \mu_\beta)/2\right\}}{\exp\left\{-(\boldsymbol{\beta}^{(curr)} - \mu_\beta)^T \sum_\beta^{-1} (\boldsymbol{\beta}^{(curr)} - \mu_\beta)/2\right\}} \,.
\end{aligned} \tag{5.4.9}$$

This step is summarized as follows.

$\boldsymbol{MH}_{\boldsymbol{\beta}}$ **step: The single MH for longitudinal coefficients**
1. Given the current state $\left(\boldsymbol{\theta}_{h_0}^{(t)}, \boldsymbol{\gamma}^{(t)}, \alpha^{(t)}, \boldsymbol{\beta}^{(t-1)}, \sigma_{\varepsilon}^{2(t-1)}, \boldsymbol{G}^{(t-1)}, \boldsymbol{b}^{(t-1)}\right)$
2. Propose new values for the vector $\boldsymbol{\beta}^{(prop)}$ from proposal distribution as in (5.4.8)
3. Calculate the acceptance probability, $r_{\boldsymbol{\beta}}$, using (5.4.9)
4. Simulate $u \sim \mathcal{U}(0,1)$:
i. If $r_{\boldsymbol{\beta}} > u$, then set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(prop)}$
ii. Else, set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$

5.4.4 $\boldsymbol{GS}_{\sigma_{\varepsilon}^2}$ and $\boldsymbol{GS}_{\boldsymbol{G}}$ steps

For the error parameter, σ_{ε}^2, and the random effects matrix $\boldsymbol{G}$, GS algorithms are used to simulate these parameters. In particular, suppose that the current parameter vector is $\left(\boldsymbol{\theta}_{h_0}^{(t)}, \boldsymbol{\gamma}^{(t)}, \alpha^{(t)}, \boldsymbol{\beta}^{(t)}, \sigma_{\varepsilon}^{2(t-1)}, \boldsymbol{G}^{(t-1)}, \boldsymbol{b}^{(t-1)}\right)$. We generate $\sigma_{\varepsilon}^{2(t)}$ directly from inverse gamma distribution in (5.3.12).

$$\sigma_{\varepsilon}^{2(t)} \sim \mathcal{IG}(\alpha^*, \beta^*)\,,$$

where

$$\alpha^* = a_0 + \frac{N}{2}\,,$$
$$\beta^* = b_0 + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta}^{(t)} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i^{(t-1)} - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i^{(t-1)}\right)^2}{2}\,.$$

Here, $\boldsymbol{b}^{(t-1)} = (\boldsymbol{v}_i^{(t-1)T}, \boldsymbol{u}_i^{(t-1)T})^T$ and $N = \sum_i^n n_i$.

For the random effects matrix, $\boldsymbol{G}$, there are two cases. When $\boldsymbol{G}$ is a non-diagonal matrix, we simulate $\boldsymbol{G}^{(t)}$ from inverse Wishart distribution in (5.3.18). In particular,

$$\boldsymbol{G}^{(t)} \sim \mathcal{I}W_q\left[n + r, \left(R + \sum_{i=1}^{n}(b_i^{(t-1)}b_i^{(t-1)T})\right)^{-1}\right]\,.$$

When the random effects matrix $\boldsymbol{G} = Diag(\sigma_1^2, ..., \sigma_q^2)$ is diagonal, we simulate $\boldsymbol{G}^{(t)}$ from the inverse gamma distribution as in (5.3.19). In particular,

$$(\sigma_1^{(t)2}, ..., \sigma_q^{(t)2}) \sim \prod_{k=1}^{q}\mathcal{IG}(a_k^*, b_k^*)\,,$$

where

$$a_k^* = a_k + \frac{n}{2}\,,$$
$$b_k^* = b_k + \frac{1}{2}\sum_{i=1}^{n} b_i^{(t-1)T} b_i^{(t-1)}\,.$$

5.4.5 $MH_{\boldsymbol{b}}$ step

The MH acceptance ratio for the random effects $\boldsymbol{b}$ has the form

$$r_{\boldsymbol{b}} = \frac{p(\boldsymbol{b}^{(prop)}|\boldsymbol{\theta}, T, \delta, y) q(\boldsymbol{b}^{(curr)}|\boldsymbol{b}^{(prop)})}{p(\boldsymbol{b}^{(curr)}|\boldsymbol{\theta}, T, \delta, y) q(\boldsymbol{b}^{(prop)}|\boldsymbol{b}^{(curr)})}\,. \tag{5.4.10}$$

Here, $\boldsymbol{b}^{(prop)}$ and $\boldsymbol{b}^{(curr)}$ are the proposed and current values of the parameter vector of the random effects $\boldsymbol{b}$ respectively. The notation $p(T, \delta, \boldsymbol{y}|.)$ is the joint likelihood function and $p(\boldsymbol{b}|\boldsymbol{\theta}, T, \delta, y)$ is the conditional posterior distribution for $\boldsymbol{b}$ as in (5.3.13). The proposal density for the coefficient vector of the longitudinal part, $q(\boldsymbol{b}^{(prop)}|\boldsymbol{b}^{(curr)})$, chosen from the independent MH (Rizopoulos, 2014) has the form

$$q(\boldsymbol{b}^{(prop)}|\boldsymbol{b}^{(curr)}) \sim \mathcal{MVN}(0, \triangle_{\boldsymbol{b}} \hat{\boldsymbol{Q}})\,, \tag{5.4.11}$$

where $\hat{\boldsymbol{Q}}$ is the asymptotic variance-covariance matrix of random effects $\boldsymbol{b}$ from the linear mixed effects model, and $\triangle_{\boldsymbol{b}}$ is the tuning parameter. Based on the conditional posterior in (5.3.13) and the proposal distribution in (5.4.11), the MH acceptance ratio for the parameter vector $\boldsymbol{b}$ has the form

$$\begin{aligned}
r_{\boldsymbol{b}} = & \frac{\prod_{i=1}^{n} \left[\exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha\left(m_i(T_i)|\boldsymbol{b}_i^{(prop)}\right)\right\}\right]^{\delta_i}}{\prod_{i=1}^{n} \left[\exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha\left(m_i(T_i)|\boldsymbol{b}_i^{(curr)}\right)\right\}\right]^{\delta_i}} \\
& \times \frac{\prod_{i=1}^{n} \exp\left(-\int_0^{T_i} h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha\left(m_i(s)|\boldsymbol{b}_i^{(prop)}\right) ds\right\}\right)}{\prod_{i=1}^{n} \exp\left(-\int_0^{T_i} h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha\left(m_i(s)|\boldsymbol{b}_i^{(curr)}\right) ds\right\}\right)} \\
& \times \frac{\prod_{i=1}^{n}\prod_{j=1}^{n_i} \exp\left\{-\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i^{(prop)} - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i^{(prop)}\right)^2 / 2\sigma_\varepsilon^2\right\}}{\prod_{i=1}^{n}\prod_{j=1}^{n_i} \exp\left\{-\left(y_i(t_{ij}) - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta} - \boldsymbol{X}_i^T(t_{ij})\boldsymbol{v}_i^{(curr)} - \boldsymbol{Z}_i^T(t_{ij})\boldsymbol{u}_i^{(curr)}\right)^2 / 2\sigma_\varepsilon^2\right\}} \\
& \times \exp\left\{-\frac{\boldsymbol{b}^{(prop)T}\boldsymbol{G}^{-1}\boldsymbol{b}^{(prop)}}{2} + \frac{\boldsymbol{b}^{(curr)T}\boldsymbol{G}^{-1}\boldsymbol{b}^{(curr)}}{2}\right\}\,.
\end{aligned} \tag{5.4.12}$$

For convenience, the sub-algorithm for the random effects $\boldsymbol{b}$ is outlined below.

	$\boldsymbol{MH_b}$ **step: The single MH for random effects**
1.	Given the current state $\left(\boldsymbol{\theta}_{h_0}^{(t)}, \boldsymbol{\gamma}^{(t)}, \alpha^{(t)}, \boldsymbol{\beta}^{(t)}, \sigma_{\varepsilon}^{2(t)}, \boldsymbol{G}^{(t)}, \boldsymbol{b}^{(t-1)}\right)$
2.	Propose new values for the vector $\boldsymbol{b}^{(prop)}$ from proposal distribution as in (5.4.11)
3.	Calculate the acceptance probability, $r_{\boldsymbol{b}}$, using (5.4.12)
4.	Simulate $u \sim \mathcal{U}(0,1)$:
	i. If $r_{\boldsymbol{b}} > u$, then set $\boldsymbol{b}^{(t)} = \boldsymbol{b}^{(prop)}$
	ii. Else, set $\boldsymbol{b}^{(t)} = \boldsymbol{b}^{(t-1)}$

5.5 Empirical results

To validate the proposed algorithms in Section 5.4, it is crucial to implement extensive simulation studies. In this section, we performed two simulation studies. In simulation study 1, linear longitudinal and survival data in Model 1 were generated for which the hazard rate at baseline had an exponential distribution and the covariance matrix of the random effects was assumed to have a non-diagonal matrix form. In simulation study 2, we simulated data from Model 2. In this model, the hazard rate at baseline had a Gompertz distribution and the covariance matrix of the random effects was assumed to have diagonal matrix form. Three knots were inserted into the longitudinal submodel.

The algorithms in Section 5.4 were applied to estimate the parameters. Based on the samples, we drew the trace plots and the density functions of the parameters in the models. Before presenting the inferences, the Geweke and Gelman diagnostics were also implemented to check the MCMC convergences. In addition, the biases and accuracy of estimates were assessed for the two models.

5.5.1 Simulation study 1

5.5.1.1 Data description

We generated the longitudinal and survival data from Model 1 in (5.2.4)

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\} \\ &= \lambda \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\}, \end{aligned} \tag{5.5.1}$$

where $h_0(t)$ has exponential distribution and the longitudinal submodel has the form

$$m_i(t) = \beta_0 + \beta_1 t + u_{i0} + u_{i1} t. \tag{5.5.2}$$

Here, $\boldsymbol{b}_i \sim N(0, \boldsymbol{D})$, $\boldsymbol{D} = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$. The data is generated in a similar way as in Section 4.1. The true values for the parameters in the joint model were $\lambda = 0.2$, $\gamma = 0.5$, $\alpha = 0.05$, $\beta_0 = 5$, $\beta_1 = 2$, $D_{11} = 1$, $D_{22} = 1$ and $D_{12} = 0.5$. The observed longitudinal value for the i^{th} subject at time point t has the form

$$y_i(t) = m_i(t) + \varepsilon_i(t),$$

where the measurement error is assumed to have normal distribution with mean 0 and standard deviation $\sigma_\varepsilon = 2$.

Based on the model in (5.5.1), we simulated the survival time for 500 subjects. The longitudinal measurements were taken once per year. For each subject, there were between 1 and 10 longitudinal measurements recorded and 1106 observations made the sample. On average, two longitudinal measurements were recorded per subject. The censoring rate was 40% of the sample.

5.5.1.2 The convergence diagnostics

The MCMC chains were created using the algorithms as described in Section 5.4. We used the Gelman and Rubin and the Geweke diagnostics to test for the convergence of the MCMC chains for all the parameters (**???**). We ran five MCMC chains of length 100,000 with a thinning of 4 for which the first 20,000 iterations were discarded as burn-in. The plots for Gelman and Rubin diagnostic are presented in Figure 5.1. A summary

of MCMC convergence of the two diagnostics are also presented in Table 5.1. In the Gelman and Rubin diagnostic tests, the potential scale reduction factors approach 1. These confirm that the MCMC chains have converged to the joint posterior distribution of the parameters. In the Geweke diagnostic tests, all of the standard Z-scores have absolute values smaller than two ($|Z| \leq 2$), which also indicates the convergence of the MCMC chains.

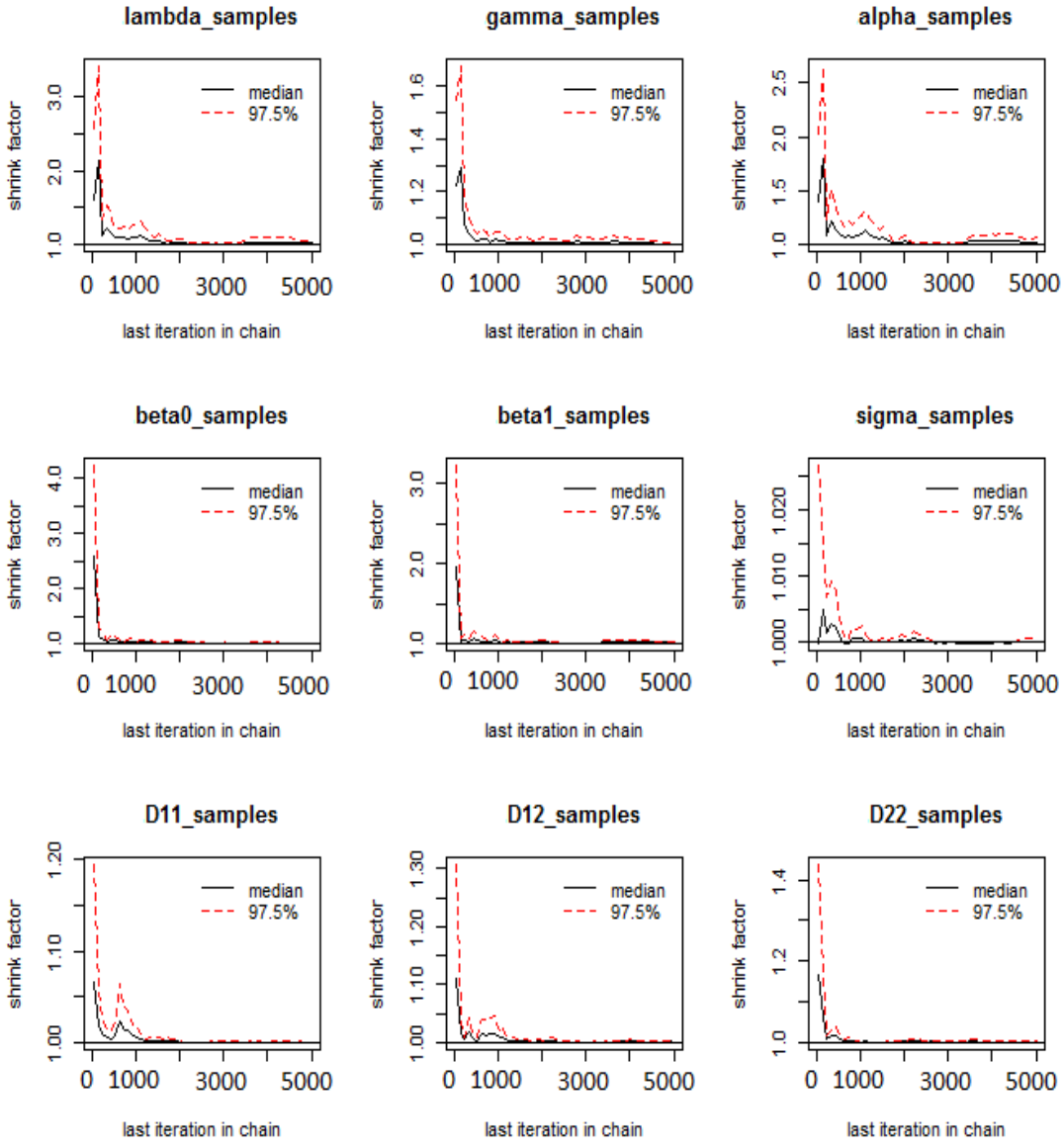

Figure 5.1: The potential rate reduction factor plots of Gelman and Rubin diagnostic for all the parameters in Model 1.

Table 5.1: Summary of MCMC convergence diagnostic tests for all the parameters in Model 1.

Gelman and Rubin diagnostic			**Geweke diagnostic**	
Potential scale reduction factors			Fraction in 1st window	0.1
	Point est.	Upper C.I.	Fraction in 2nd window	0.5
λ	1.02	1.06	λ	-0.625
γ	1.00	1.01	γ	0.479
α	1.03	1.07	α	0.377
β_0	1.01	1.01	β_0	0.304
β_1	1.00	1.03	β_1	-0.653
σ^2_ϵ	1.00	1.00	σ^2_ϵ	0.250
D_{11}	1.00	1.00	D_{11}	0.212
D_{12}	1.00	1.00	D_{12}	0.582
D_{22}	1.00	1.00	D_{22}	0.555
Multivariable psrf	1.02			

5.5.1.3 Parameter estimation

The algorithms as described in Section 5.4 were used to estimate the parameters in Model 1. We ran 100,000 iterations of the algorithm. The thinning was applied to reduce the autocorelation in the samples. This means that we only kept the values from the samples at particular steps. The convergence diagnostics tests confirmed the convergence of each simulated sample to the stationary target distribution in the previous section. We kept the last 5,000 iterations for making inferences. The traces of the parameter samples and the posterior distributions are presented in Figures 5.2, 5.3 and 5.4 respectively. The thick lines in the posterior distribution indicate the positions of the true values for each parameter.

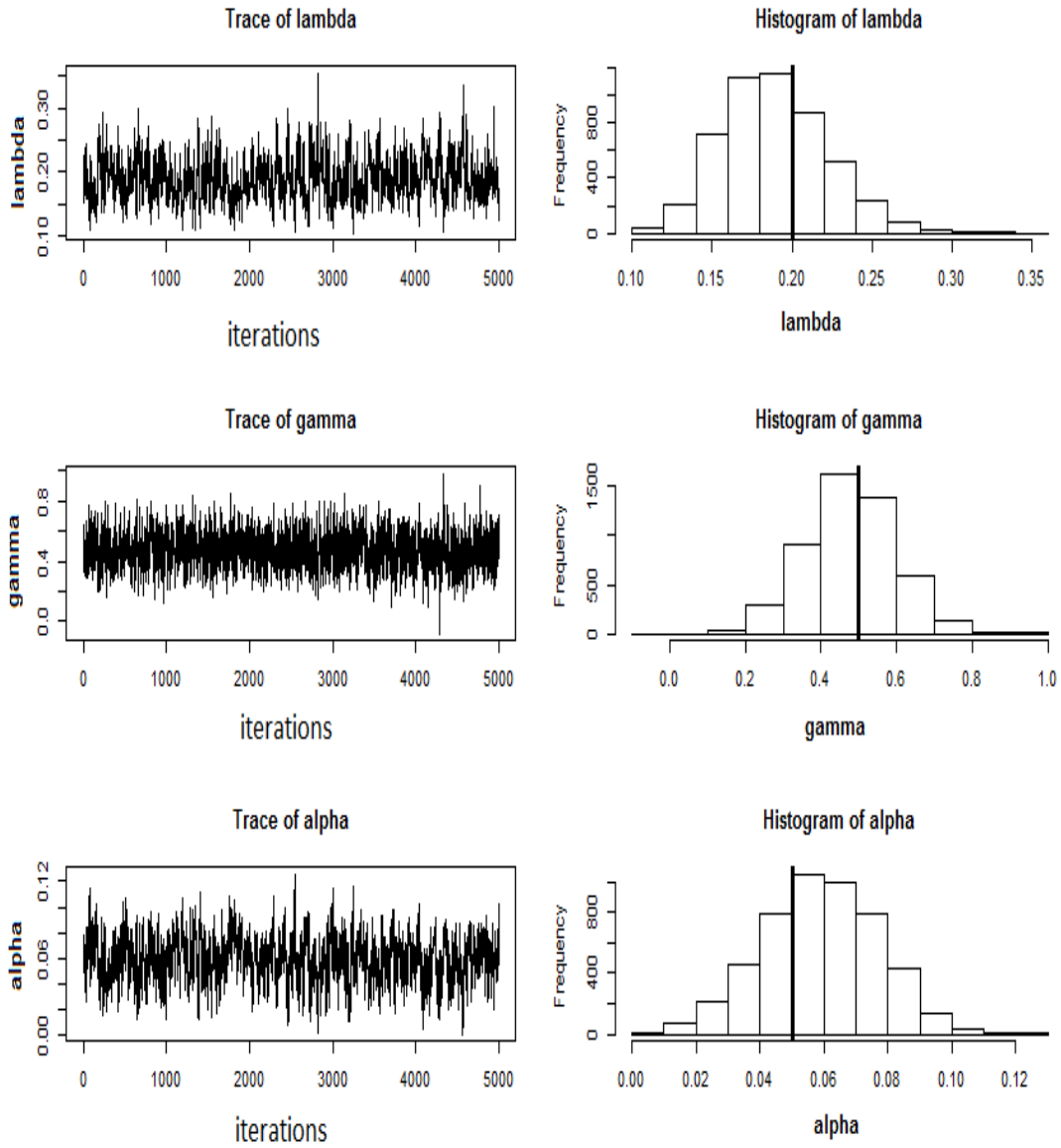

Figure 5.2: MCMC traces and posterior distribution plots for the parameters λ, γ and α in Model 1. The thick line indicates the position of the true value.

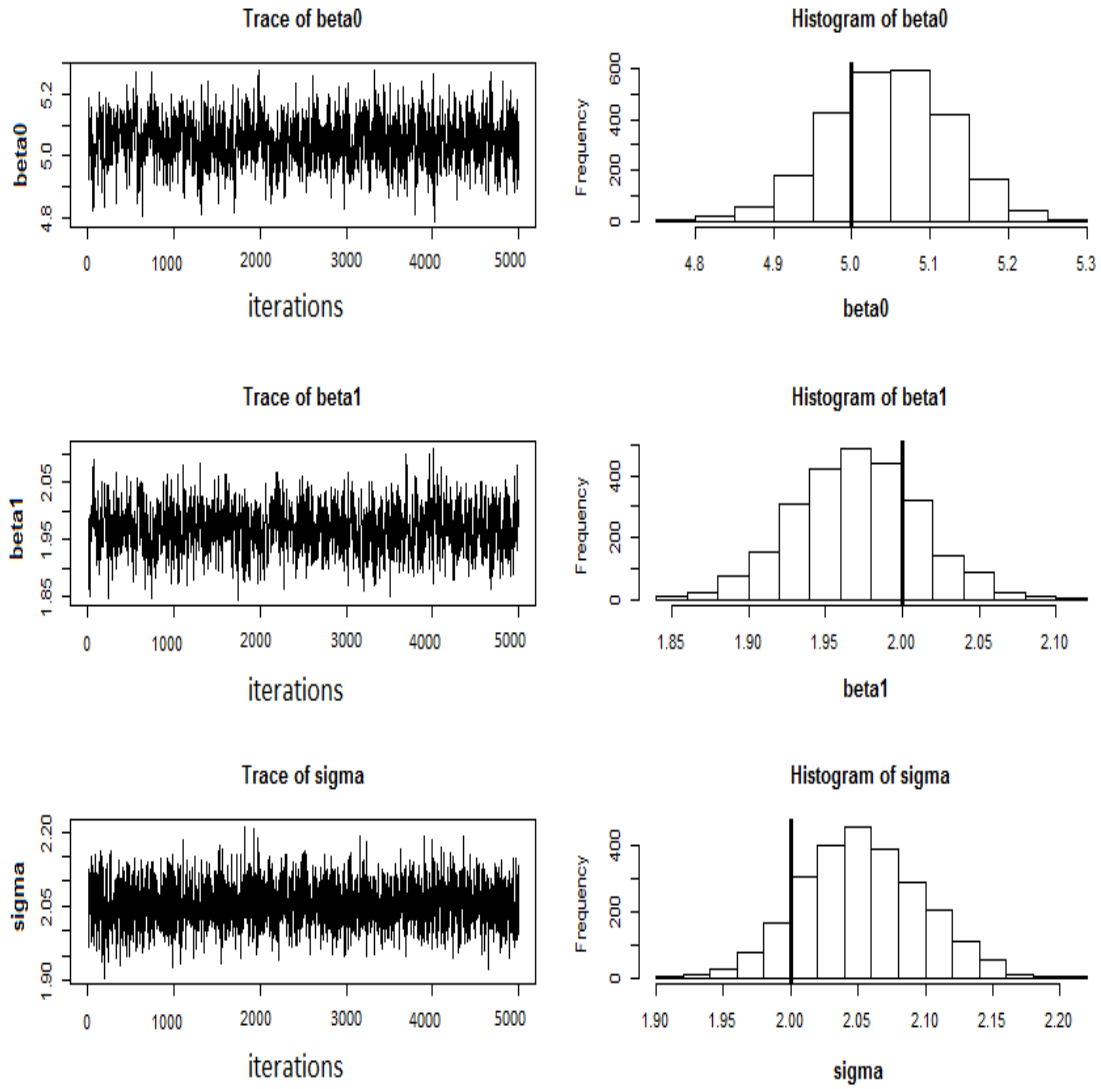

Figure 5.3: MCMC traces and posterior distribution plots for the parameters β_0, β_1 and σ in Model 1. The thick line indicates the position of the true value.

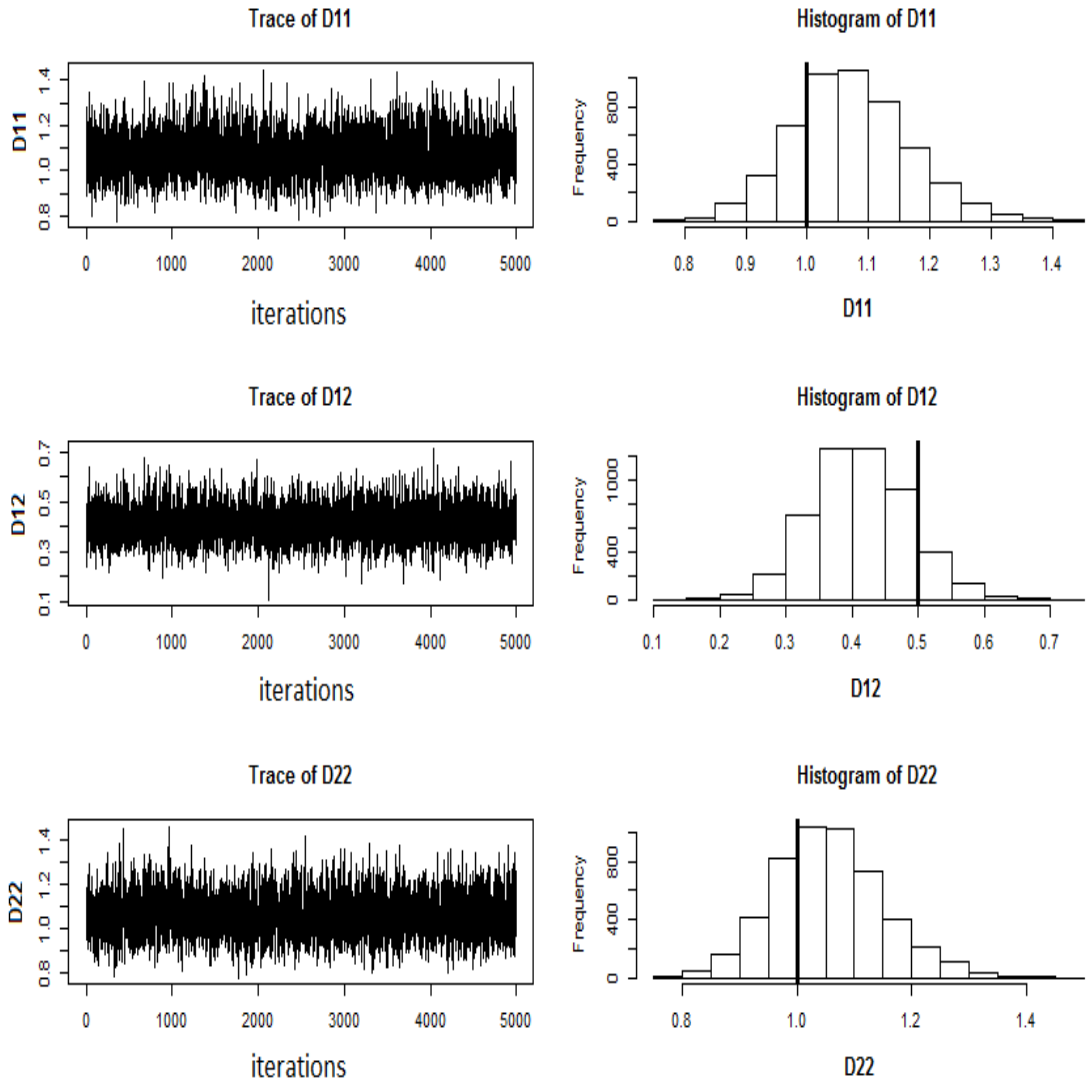

Figure 5.4: MCMC traces and posterior distribution plots for the parameters D_{11}, D_{12} andD_{22} in Model 1. The thick line indicates the position of the true value.

The acceptance rates for the MH algorithms were 33.67% for λ, 49.11% for γ, 31.66% for α, 44.69% for β_0 and 49.46% for β_1. For all parameters in Figures 5.2, 5.3 and 5.4, the samples mix well. Moreover, the trace plots have shown the stability of the MCMC chains for each parameter and also demonstrate the convergence of the samples to the posterior distributions. In general, the marginal posterior distributions have an unimodal type so that one can infer information about the centre and the spread of each parameter.

The autocorrelation function (ACF) plots for all the parameters in Model 1 are presented in Figure 5.5. In the figure, we can see that the ACF plots for all of the parameters decay

relatively quickly to 0. These plots also show that the chains for all the parameters are mixing well and the subsequent samples in the chains are independent as the lag increases. Especially, the ACF plots for γ, β_0, β_1, σ_ε^2 , D_{11}, D_{12} and D_{22} cut off at around lag 3 and tend to towards zero until about 20 or 25 lags, while the ACF plots for λ and α decrease exponentially to 0.

There are slow decaying autocorrelations for the parameters λ and α as depicted in Figure 5.5. This might be due to the impact of prior distributions on the autocorrelation. We use flat prior distributions for both parameters in this analysis. In joint models, the baseline hazard rate function is unspecified and the association parameter, α, has little prior information. Therefore, a prior sensitivity analysis is performed for these parameters in Section 5.6.

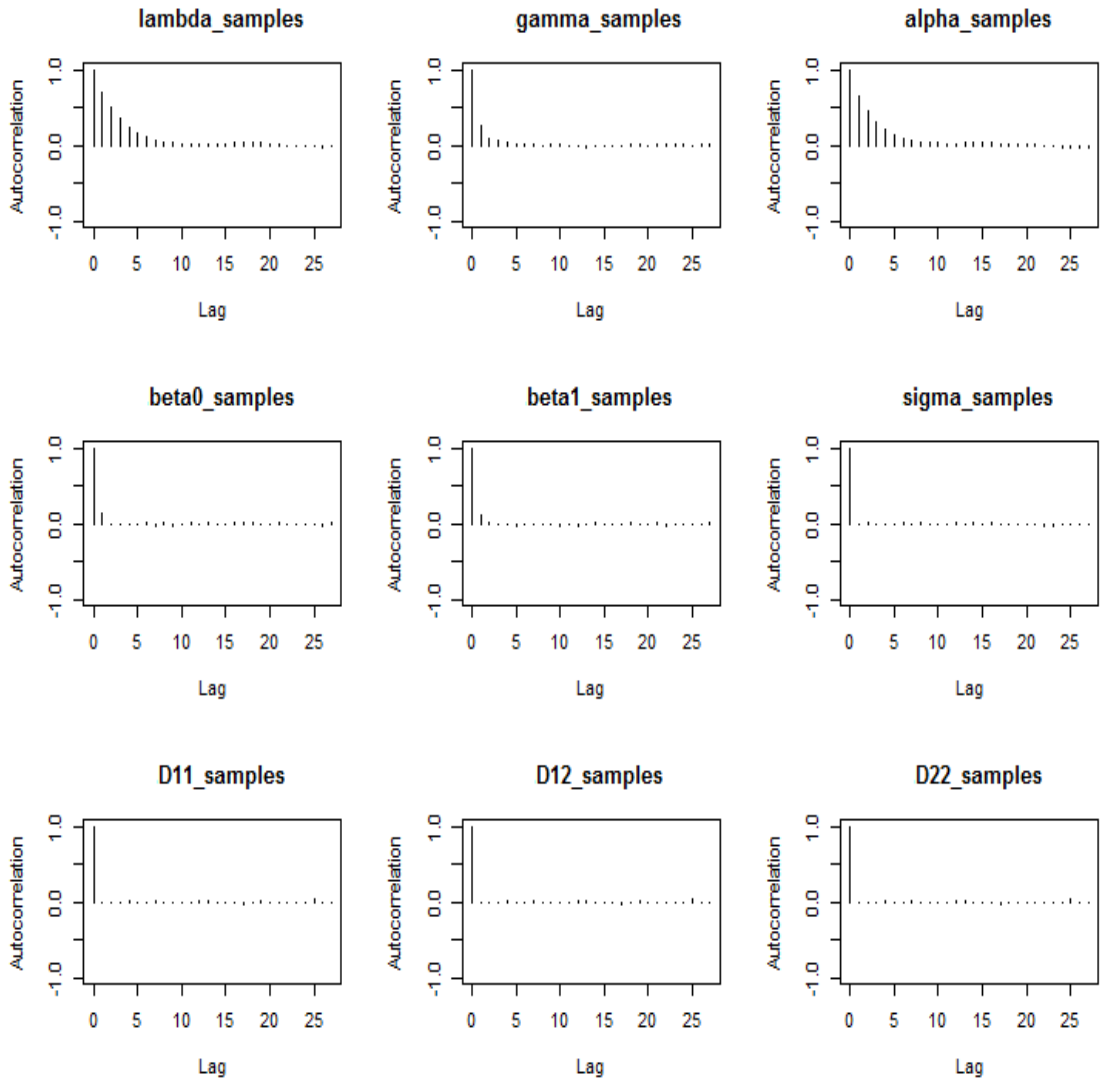

Figure 5.5: ACF plots for all the parameters in Model 1.

In order to evaluate the accuracy of the estimates, we generated 50 independent datasets from Model 1. The means of the estimates, 95% credible intervals (CrIs) and CrI performance are presented in Table 5.2. Here, the coverage performance is the percentage of true values that lie in the CrIs. In this table, the point estimates are reasonably close to the true values for all the parameters. Moreover, the true values are also within their 95% CrIs.

Table 5.2: Summary statistics for parameter estimation of the simulated data of the models in (5.5.1) and (5.5.2).

Parameter	True value	Mean	SD	95% CrI
λ	0.2	0.235	0.046	[0.159;0.469]
γ	0.5	0.473	0.103	[0.218;0.674]
α	0.05	0.035	0.018	[0.002;0.075]
β_0	5	5.101	0.154	[4.782;5.414]
β_1	2	1.876	0.228	[1.501;2.118]
σ	2	2.062	0.222	[1.898;2.177]
D_{11}	1	0.911	0.454	[0.770;1.384]
D_{12}	0.5	0.394	0.184	[0.271;0.761]
D_{22}	1	0.843	0.743	[0.727;1.334]

5.5.2 Simulation study 2

5.5.2.1 Data description

We generated the longitudinal and survival data from Model 2 (5.2.6)

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\} \\ &= \lambda_1 \exp(\lambda_2 t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + \alpha m_i(t)\right\} . \end{aligned} \tag{5.5.3}$$

Here, $h_0(t)$ has Gompertz distribution and the longitudinal submodel has the form

$$m_i(t) = \beta_0 + \beta_1 t + u_{i0} + \sum_{k=1}^{3} u_{ik}(t - \mathcal{K}_k)_+^p . \tag{5.5.4}$$

Recall that $\boldsymbol{b}_i = (u_{i0}, u_{i1}, u_{i2}, u_{i3})^T$ is a vector of random effects for the i^{th} subject. We assume that the random effects, $\boldsymbol{b}_i$, follow a normal distribution with mean 0 and the covariance matrix $\boldsymbol{D} = Diag(D_{11}, D_{22}, D_{33}, D_{44})$. The steps of generating the data are as presented in Section 4.1. The true values for the parameters were $\lambda_1 = 0.1$, $\lambda_2 = 0.5$ $\gamma = 0.5$, $\alpha = 0.05$, $\beta_0 = 5$, $\beta_1 = 2$, $D_{11} = 4$, $D_{22} = 2$, $D_{33} = 2$ and $D_{44} = 2$. The observed longitudinal value for the i^{th} subject at time point t has the form

$$y_i(t) = m_i(t) + \varepsilon_i(t) ,$$

where the measurement error is assumed to have normal distribution with mean 0 and standard deviation $\sigma_\varepsilon = 2$.

Based on the model in (5.5.3), we simulated the survival time for 500 subjects. The longitudinal measurements were taken each year. For each subject, there were 1-5 longitudinal measurements recorded with a total of 1387 observations for the sample. On average, three longitudinal measurements were recorded per subject. The censoring rate was 35% of the sample.

5.5.2.2 The convergence diagnostics

Before presenting the inferences for the parameters in Model 2, we report the convergence diagnostic tests for the MCMC chains generated by the algorithms in Section 5.4. The Gelman and Rubin and the Geweke convergence diagnostics were used to test for all the parameters in Model 2. In the Gelman and Rubin test, five MCMC chains for each parameter were simulated with 100,000 iterations. The thinning is 4 and the first 20,000 iterations were discarded as burn-in. The plots for the Gelman and Rubin diagnostic are presented in Figures 5.6 and 5.7. A summary of the two diagnostics are also presented in Table 5.3.

In the Gelman and Rubin diagnostic tests, the potential scale reduction factors for all the parameters in Model 2 reduce quickly. They are very close to 1 when the iteration is large. The multivariable potential scale reduction factors is 1.02. As a results of these, the MCMC chains are diagnosed to converge to the conditional posterior distributions of the parameters. The convergence of the MCMC chains are confirmed again in the Geweke diagnostic tests, where the standard Z-scores all have absolute values smaller than two ($|Z| \leq 2$).

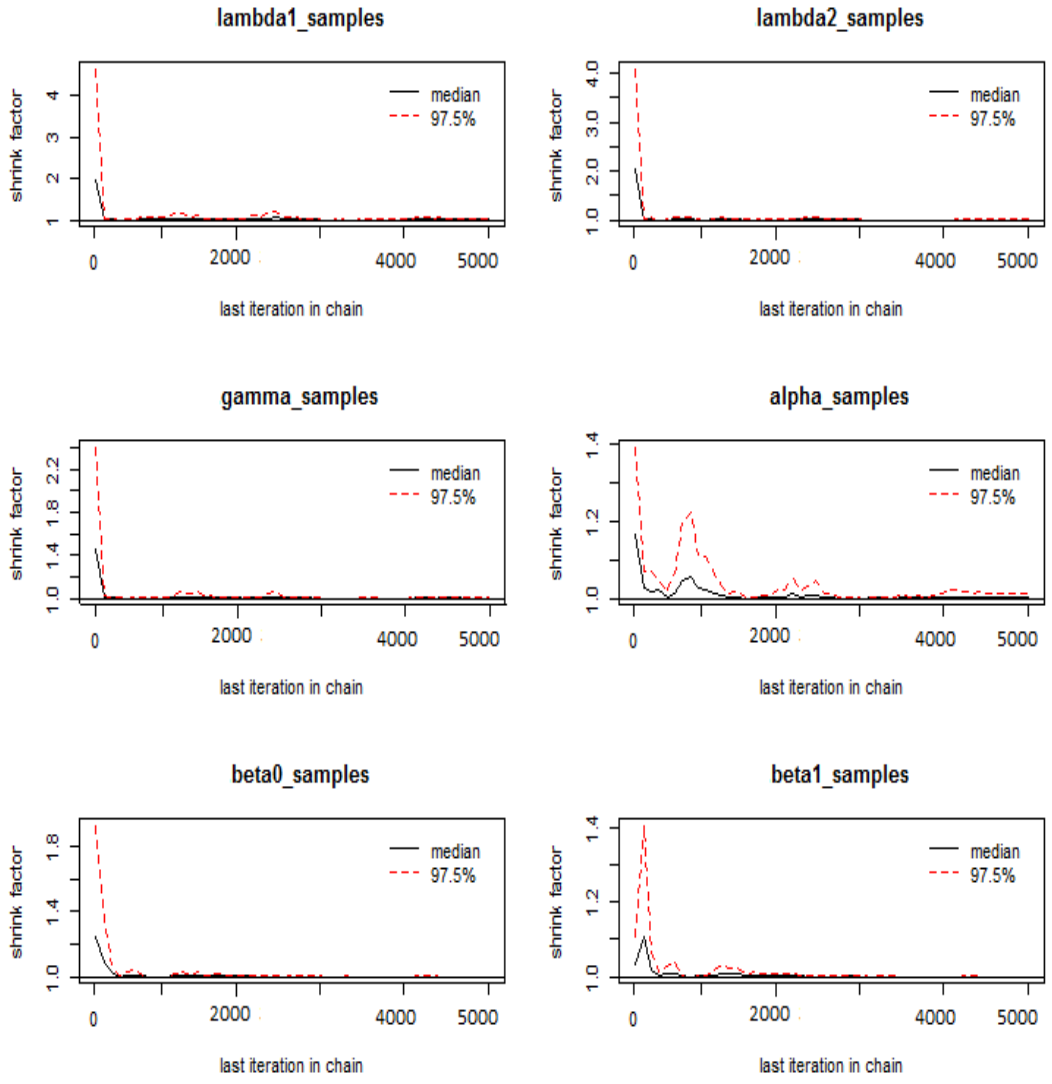

Figure 5.6: The potential rate reduction factor plots from Gelman and Rubin diagnostic for the parameters λ_1, λ_2, γ, α, β_1 and β_2 in Model 2.

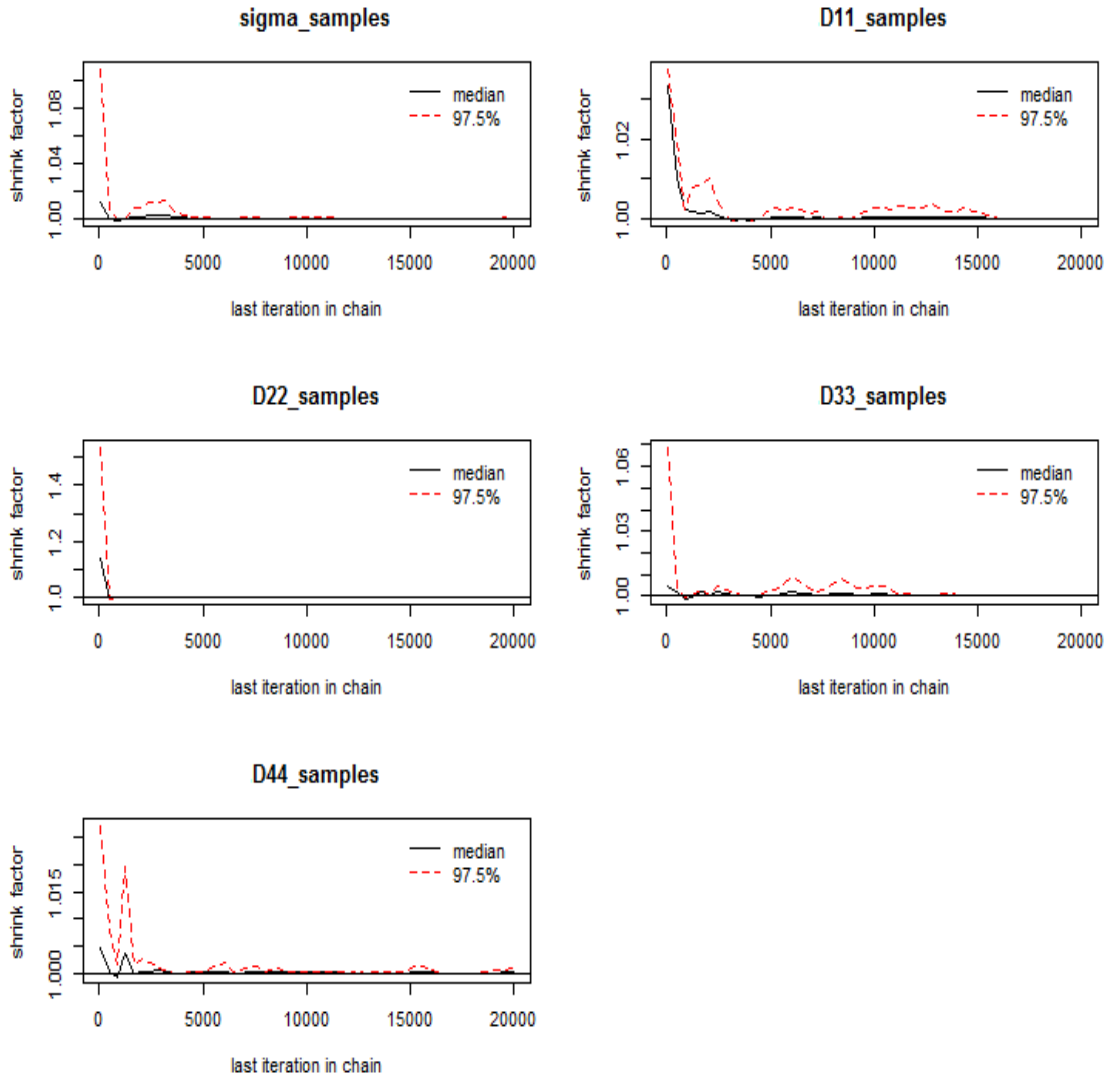

Figure 5.7: The potential rate reduction factor plots from Gelman and Rubin diagnostic for the parameters σ_ε, D_{11}, D_{22}, D_{33} and D_{44} in Model 2.

Table 5.3: Summary of MCMC convergence diagnostic tests for all the paramters in Model 2.

Gelman and Rubin diagnostic			**Geweke diagnostic**	
Potential scale reduction factors			Fraction in 1st window	0.1
	Point est.	Upper C.I.	Fraction in 2nd window	0.5
λ_1	1.02	1.06	λ_1	-0.296
λ_2	1.02	1.06	λ_2	1.536
γ	1.00	1.01	γ	0.510
α	1.02	1.06	α	-0.982
β_0	1.00	1.01	β_0	0.032
β_1	1.00	1.00	β_1	-0.270
σ^2_ϵ	1.0	1.00	σ^2_ϵ	0.167
D_{11}	1.00	1.00	D_{11}	-0.555
D_{22}	1.0	1.00	D_{22}	0.522
D_{33}	1.0	1.00	D_{33}	-0.208
D_{44}	1.00	1.00	D_{44}	0.478
Multivariable psrf	1.02			

5.5.2.3 Parameter estimation

The algorithms as described in Section 5.4 were implemented to estimate the parameters in Model 2. We also ran 100,000 iterations of the algorithm and kept the last 5,000 iterations for making inferences. The traces of the parameter samples and the posterior distributions are presented in Figures 5.8, 5.9, 5.10 and 5.11. The thick lines in the density functions show the true values of the parameters.

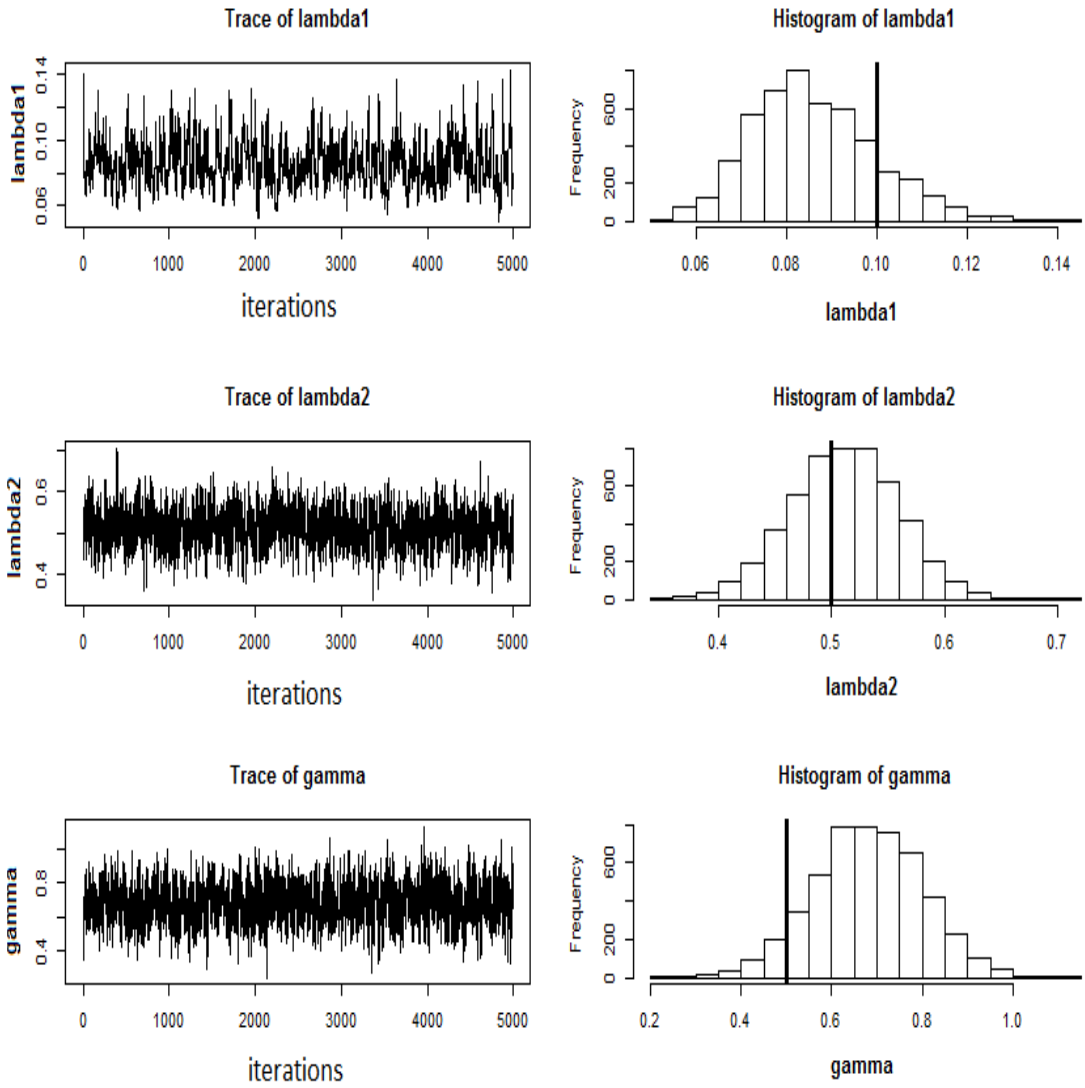

Figure 5.8: MCMC traces and posterior distribution plots for the parameters λ_1, λ_2, and γ in Model 2. The thick line indicates the position of the true value.

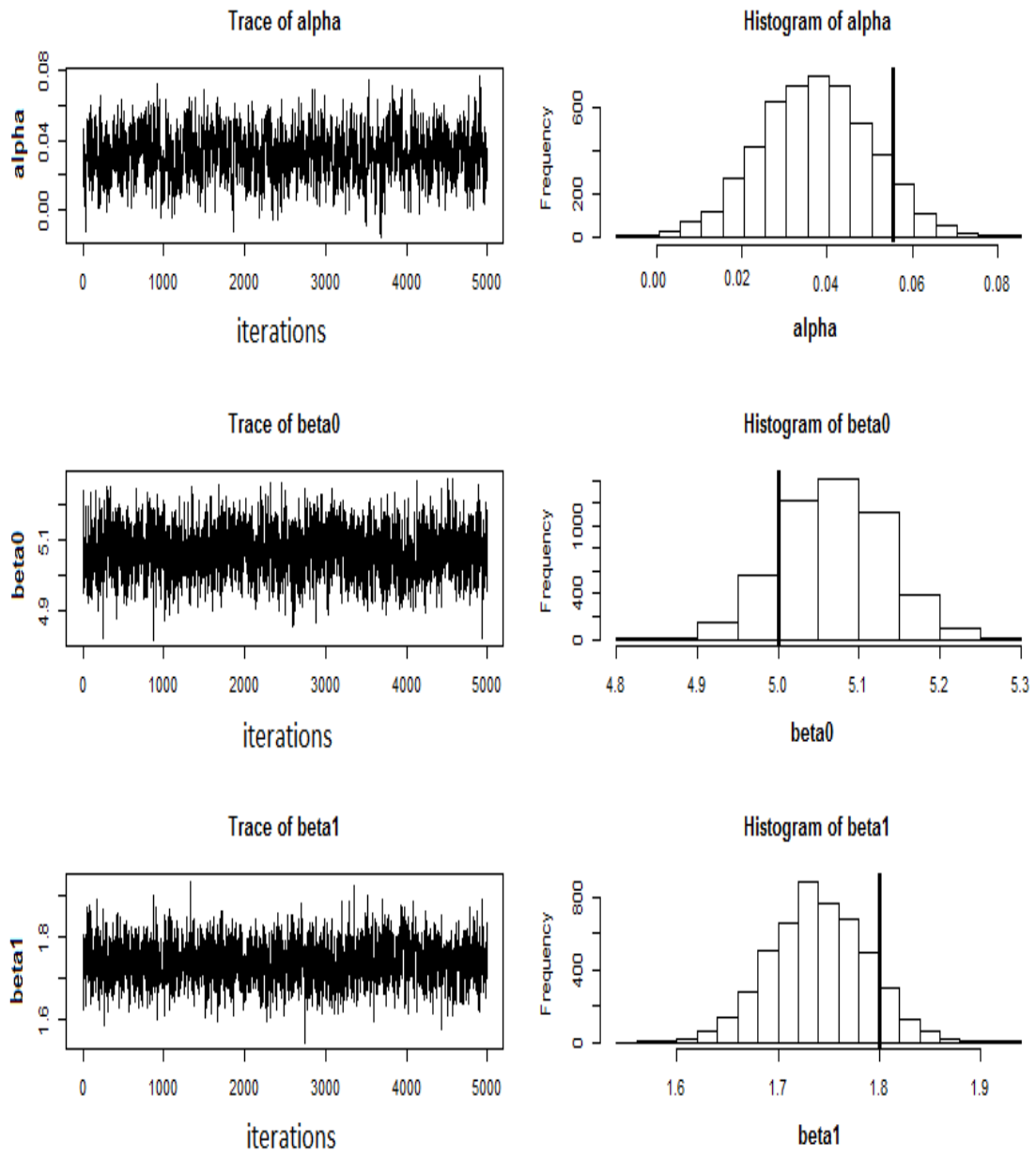

Figure 5.9: MCMC traces and posterior distribution plots for the parameters α, β_0 and β_1 in Model 2. The thick line indicates the position of the true value.

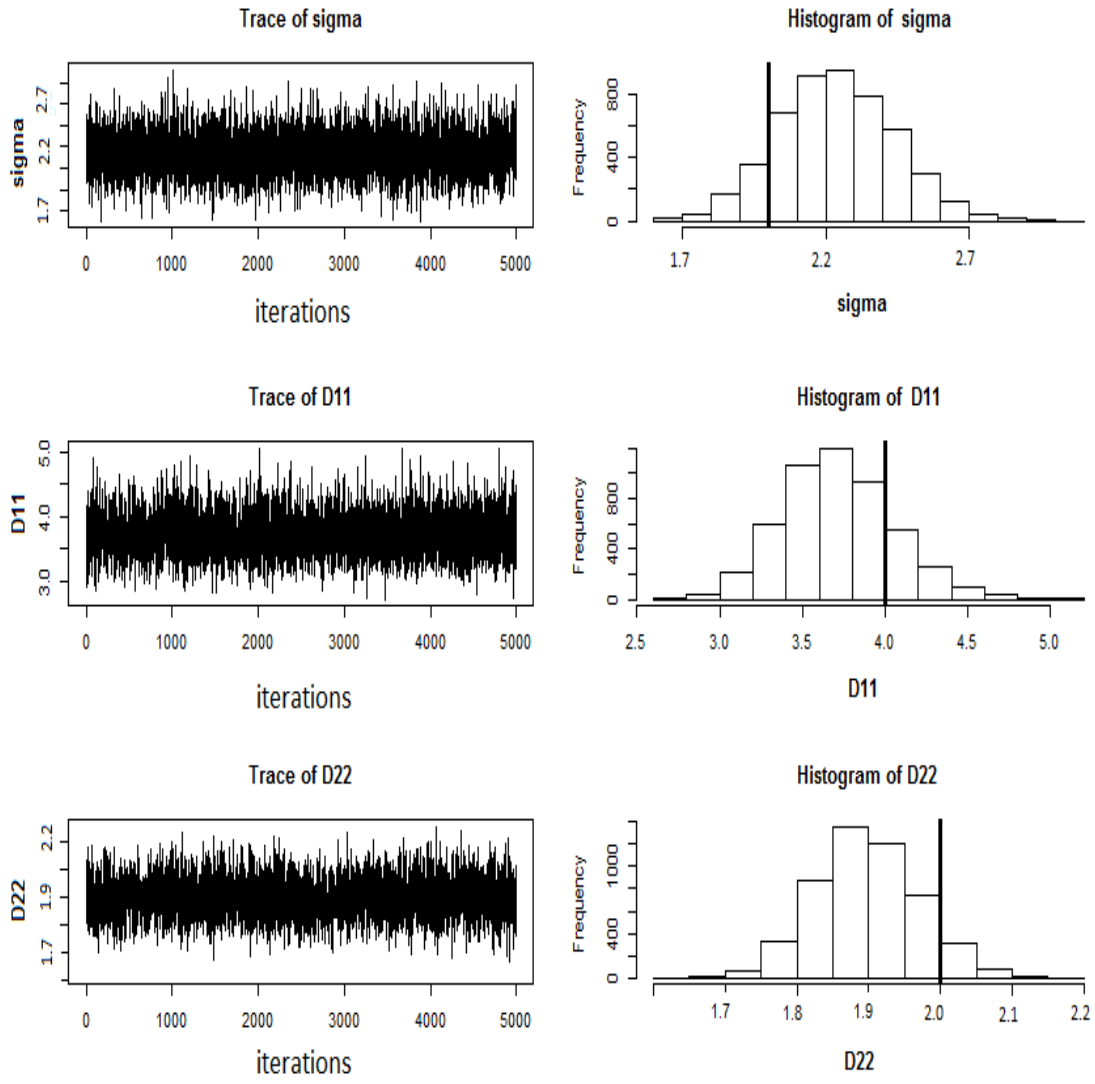

Figure 5.10: MCMC traces and posterior distribution plots for the parameters σ_ε^2 , D_{11} and D_{22} in Model 2. The thick line indicates the position of the true value.

From Figures 5.8 and 5.9, the density functions of these samples show relatively unimodal distributions. The true values of these parameters are around the center of the distributions. From Figures 5.10 and 5.11, the samples were generated from GS algorithms. Therefore, the acceptance rate is always 1. The density functions of these samples are unimodal distributions in which the true values of the parameters are contained within the three standard deviation range. For all parameters, the samples are distributed around the mean and the trace plots are stable over time. These visually show that the MCMC chains converge to the target posterior distributions.

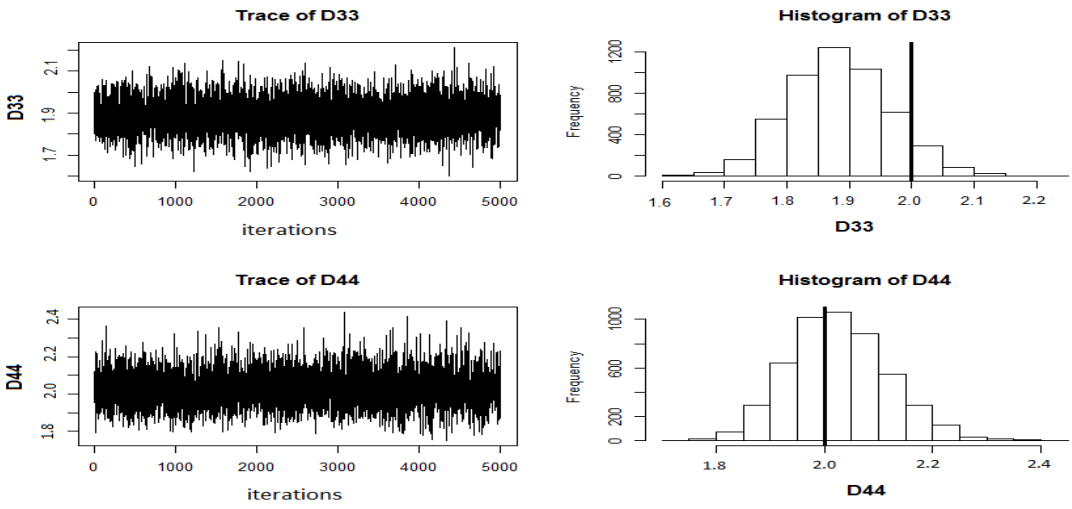

Figure 5.11: MCMC traces and posterior distribution plots for the parameters D_{33} and D_{44} in Model 2. The thick line indicates the position of the true value.

The ACF plots for all of the parameters in Model 2 are presented in Figures 5.12 and 5.13. From the figures, we can see that in general, the ACF plots for all of the parameters decrease quickly to 0. In particular, the ACF plots for the parameters of the hazard rate at baseline, the coefficients of the longitudinal submodel and the coefficients of the survival submodel decrease exponentially and go to 0 at around lag 5. The ACF plots for σ_ε^2 , D_{11}, D_{22}, D_{33} and D_{44} cut off very quickly at around lag 3 and tend towards zero. These plots show that the chains for all of the parameters mix well and the subsequent samples in the chains are independent as the lag increases.

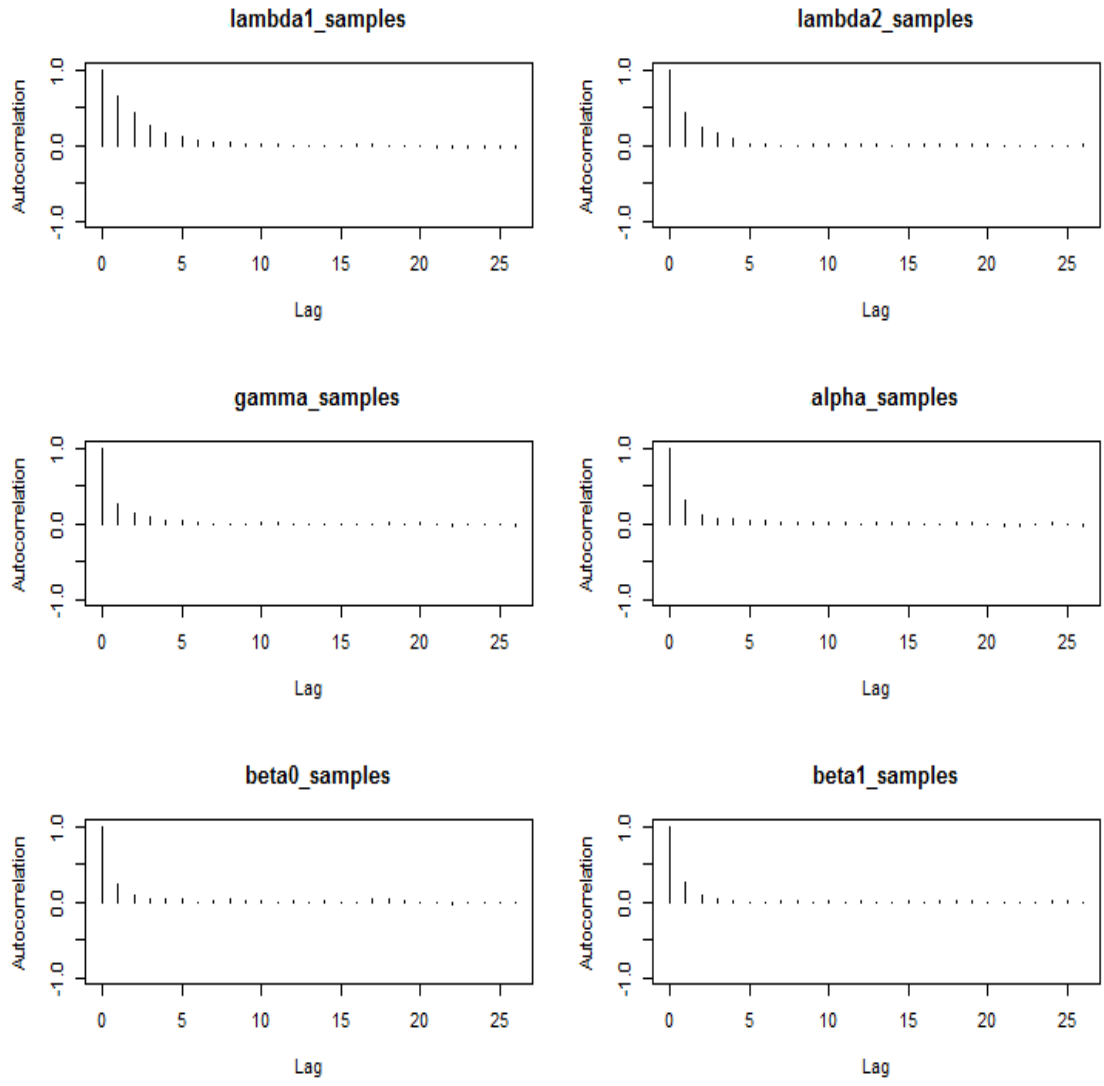

Figure 5.12: ACF plots for the parameters λ_1, λ_2, γ, α, β_1 and β_2 in Model 2.

In order to evaluate the accuracy of the estimates, thirty independent datasets were generate from Model 2. The means of the estimates, 95% CrIs and CrI performance are presented in Table 5.4. In this table, the estimates for the parameters of the hazard rate at baseline,$\lambda_1$and λ_2, are close to the true values. Moreover, the true values are within their 95% CrIs and the CrI performances are more than 93%. Similarly, the estimates for the coefficient parameters of the survival part, (γ, α), the coefficient parameters for the longitudinal part, (β_0, β_1) and error measurement, σ_ε^2, are all close to the true values. The CrIs contain the true values of the parameters and the CrI performances are really good, especially for the parameters λ_1 and γ which have coverage percentages more than

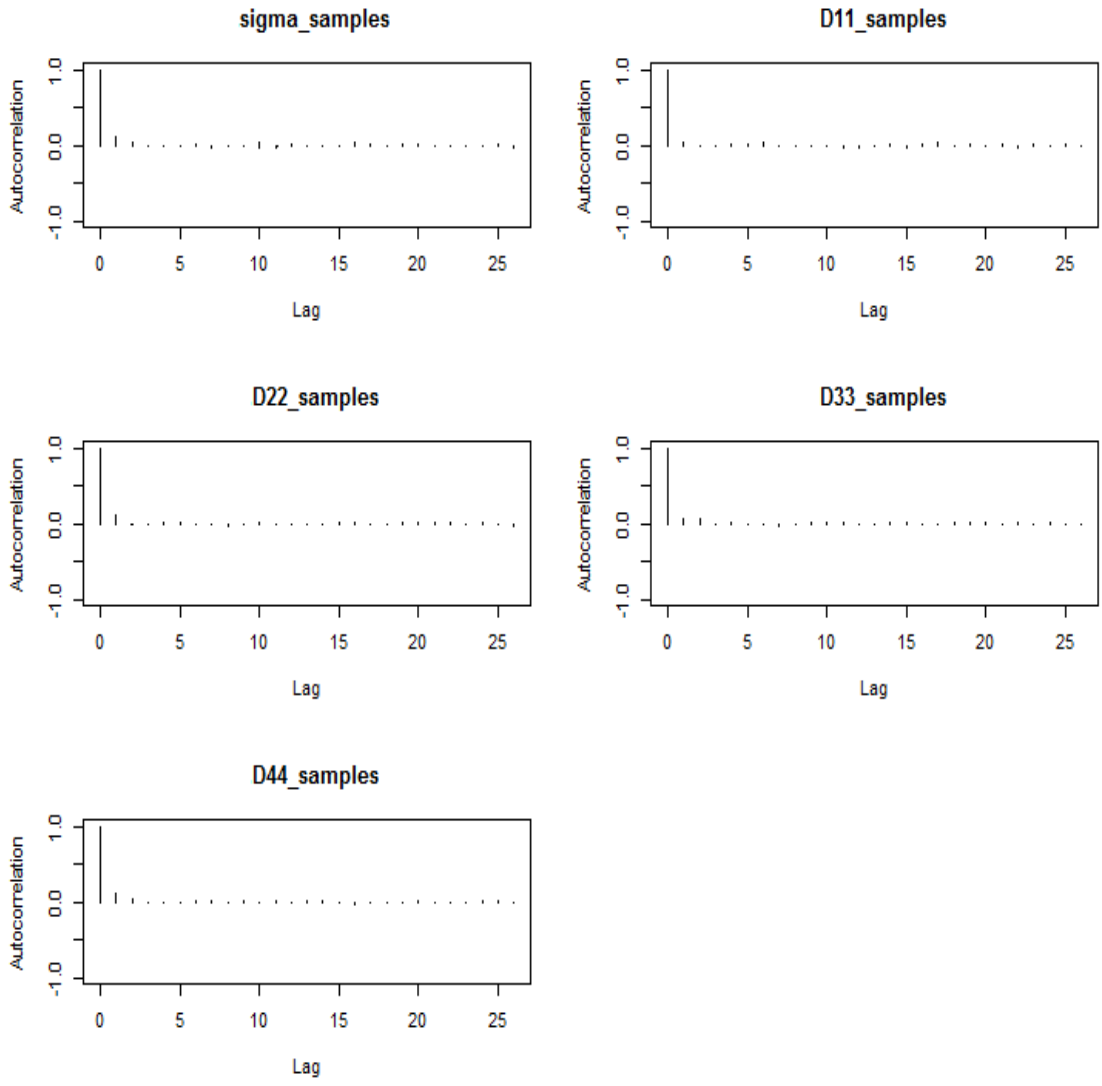

Figure 5.13: ACF plots for the parameters σ_ϵ^2, D_{11}, D_{22}, D_{33} and β_2 in Model 2.

Table 5.4: Summary statistics for parameter estimation of the simulated data of the model in (5.5.3) and (5.5.4).

Parameter	True value	Mean	SD	95% CrI	CrI performance
λ_1	0.1	0.115	0.017	[0.008;0.146]	96.67%
λ_2	0.5	0.514	0.049	[0.415;0.614]	93.33%
γ	0.5	0.467	0.112	[0.250;0.697]	96.67%
α	0.05	0.034	0.012	[0.012;0.069]	93.33%
β_0	5	5.015	0.319	[4.496;5.334]	90%
β_1	2	1.983	0.356	[1.645;2.320]	93.33%
σ	2	2.251	0.176	[1.756;2.466]	96.67%
D_{11}	4	4.082	0.523	[3.458;4.898]	92%
D_{22}	2	2.369	0.383	[1.981;2.798]	90%
D_{33}	2	2.502	0.243	[1.525;3.578]	90%
D_{44}	2	1.932	0.807	[1.727;3.558]	86.67%

95%.

For the parameters of the random effects, D_{11} , D_{22} D_{33} and D_{44}, the estimates are relatively close to the true value. However, the 95% CrIs have large ranges and the CrI performances for these parameters are lowest among all the parameters of Model 2. The reason for this is that the dimension of random effects is larger in this simulation setting than in the simulation study 1. According to Robert and Casella (2004), the independent MH algorithms have a limitation in high-dimensional models where the forms of the conditional posterior distributions are often complicated. This can affect the accuracy of estimation and coverage performance. Moreover, the missing data affects the estimates of the random effects when we put many knots into the model (Rizopoulos, 2012).

5.6 Prior sensitivity analysis

In this section, we conduct a prior sensitivity analysis for the parameter of the hazard rate at baseline, λ, and the association parameter between longitudinal data and survival data, α, of Model 1. The hazard rate at baseline, $h_0(t)$, is unspecified in the joint model. Therefore, we have only a little information about the prior distribution of this parameter. In addition, the hazard rate at baseline has a direct influence on the hazard rate function in

Table 5.5: Summary of prior type for the baseline hazard rate, λ, and the association parameter, α.

Prior type	λ	α	Detail
I	$\mathcal{N}(\tilde{\mu}_\lambda, 49\tilde{\sigma}^2_\lambda)$	$\mathcal{N}(\tilde{\mu}_\alpha, 49\tilde{\sigma}^2_\alpha)$	Relatively uninformative prior on λ and relatively uninformative prior on α
II	$\mathcal{N}(\tilde{\mu}_\lambda, 100\tilde{\sigma}^2_\lambda)$	$\mathcal{N}(\tilde{\mu}_\alpha, 49\tilde{\sigma}^2_\alpha)$	Flat prior on λ and relatively uninformative prior on α
III	$\mathcal{N}(\tilde{\mu}_\lambda, 49\tilde{\sigma}^2_\lambda)$	$\mathcal{N}(\tilde{\mu}_\alpha, 100\tilde{\sigma}^2_\alpha)$	Relatively uninformative prior on λ and flat prior on α
IV	$\mathcal{N}(\tilde{\mu}_\lambda, 100\tilde{\sigma}^2_\lambda)$	$\mathcal{N}(\tilde{\mu}_\alpha, 100\tilde{\sigma}^2_\alpha)$	Flat prior on λ and flat prior on α

the joint models. In addition, the association parameter is the most important parameter for evaluating the link between longitudinal and survival data.

To choose the prior distribution for these parameters, we take full advantage of the ordinary two-stage approach. In particular, the separate estimates from survival data are used to define the mean and variance in the prior distributions of λ. We take the estimated means, $\tilde{\mu}_\lambda$, and estimated variances, $\tilde{\sigma}^2_\lambda$, of parameter λ from the ordinary two-stage approach. Note that these estimates are more biased than the full likelihood approach as proved in Chapter 4 and in Sweeting and Thompson (2011). We choose the prior distribution for λ having a normal distribution $\mathcal{N}(\mu_\lambda, \sigma^2_\lambda)$. The prior mean, μ_λ, is the estimated mean, $\tilde{\mu}_\lambda$, in the two-stage approach. The prior variance, σ^2_λ, is chosen large enough to have a relatively uninformative prior on λ. Here, for the sake of convinience, there are two chosen prior variances which are 49 and 100 times the estimated variance, $\tilde{\sigma}^2_\lambda$, in the ordinary two-stage approach.

In a similar way, we employ the estimates from the coxph function for α and use them to define the parameters in the prior distribution of α. The normal distribution $\mathcal{N}(\mu_\alpha, \sigma^2_\alpha)$ is chosen for the prior distribution of α. The mean, μ_α, is chosen from the estimated value, $\tilde{\mu}_\alpha$. The variance, σ^2_α, is chosen larger than the estimated variance, $\tilde{\sigma}^2_\alpha$, by 49 and 100 times. The prior distributions for λ and α are labeled from I to IV respectively and summarized in Table 5.5.

In Table 5.7, the estimates of posterior means, standard deviation and 95% CrIs are presented for priors I, II and III based on 50 independent samples. Note that, the estimates for prior IV were presented previously in simulation study 1, Table 5.2. In general, the

four priors performed very well and are comparable to each other. The estimated means of all parameters are very close to the true values with reasonably small standard deviations. In addition, the 95% CrIs include the true values for the four priors. This indicates that the statistical inferences are insensitive to different types of priors of λ and α.

The coverage performances are presented in Table 5.6 for the four priors. The coverage percentages for each parameter are comparable between the four priors. This again confirms that different types of prior distributions of λ and α do not evidently affect the Bayesian inferences. In particular, for the survival parameters, λ, γ and α, the coverage percentage is very high around 96% on average for the four priors. This also indicates the accuracy of the estimates for these parameters through the algorithm in Section 5.4. For longitudinal parameters, β_0, $\beta_1$and σ_ε^2, the coverage is around 94% on average and 92% on average for random effects, D_{11}, D_{12} and D_{22}.

Table 5.6: Coverage performance of Model 1 for different prior types.

	Performance			
Parameter	Prior I	Prior II	Prior III	Prior IV
λ	98%	98%	100%	98%
γ	96%	100%	96%	98%
α	94%	96%	98%	100%
β_0	94%	94%	96%	96%
β_1	94%	94%	94%	92%
σ	92%	94%	94%	94%
D_{11}	96%	94%	94%	92%
D_{12}	92%	94%	92%	92%
D_{22}	90%	90%	92%	90%

Table 5.7: Summary statistics for parameter estimation of the simulated data of Model 1 for different prior types.

Parameter	True	Prior I			Prior II		
		Mean	SD	95% CrI	Mean	SD	95% CrI
λ	0.2	0.217	0.032	[0.156;0.289]	0.215	0.034	[0.161;0.269]
γ	0.5	0.482	0.105	[0.258;0.710]	0.470	0.102	[0.247;0.696]
α	0.05	0.046	0.014	[0.014;0.077]	0.046	0.016	[0.007;0.071]
β_0	5	5.084	0.185	[0.934;5.233]	5.062	0.154	[4.911;5.212]
β_1	2	1.844	0.224	[1.758;2.193]	1.843	0.173	[1.789;2.196]
σ	2	2.071	0.331	[1.985;2.161]	2.069	0.207	[1.988;2.166]
D_{11}	1	0.991	0.339	[0.829;1.175]	0.860	0.389	[0.700;1.199]
D_{12}	0.5	0.478	0.210	[0.350;0.619]	0.518	0.196	[0.319;0.558]
D_{22}	1	0.933	0.296	[0.780;1.107]	0.905	0.320	[0.738;1.148]
Parameter	True	Prior III			Prior IV		
		Mean	SD	95% CrI	Mean	SD	95% CrI
λ	0.2	0.213	0.057	[0.154;0.284]	0.235	0.046	[0.159;0.469]
γ	0.5	0.507	0.137	[0.282;0.733]	0.473	0.103	[0.218;0.674]
α	0.05	0.046	0.013	[0.015;0.077]	0.035	0.018	[0.002;0.075]
β_0	5	5.035	0.133	[4.886;5.183]	5.101	0.154	[4.782;5.414]
β_1	2	1.871	0.196	[1.787;2.195]	1.876	0.228	[1.501;2.118]
σ	2	2.089	0.217	[1.985;2.133]	2.062	0.222	[1.898;2.177]
D_{11}	1	0.925	0.394	[0.774;1.185]	0.911	0.454	[0.770;1.384]
D_{12}	0.5	0.409	0.182	[0.289;0.541]	0.394	0.184	[0.271;0.761]
D_{22}	1	0.904	0.307	[0.756;1.173]	0.843	0.743	[0.727;1.334]

5.7 Case study

In this section, we consider liver cirrhosis data. In the liver cirrhosis dataset, there are 488 patients who are diagnosed with liver cirrhosis. The patients are divided randomly into two groups. One group received prednisone and another receive a placebo. This study was conducted from 1962 to 1974 in Copenhagen (Andersen et al., 1993). The prothrombin index was recorded at 3, 6, 12 months and yearly thereafter. By the end of the study, 150 patients receiving the prednisone treatment died (63.3%) and 142 patients receiving the placebo treatment died (56.6%). Resulting in a censoring rate at about 38.73%. There

were 2968 longitudinal responses recorded, accounting for 64.22% of missing responses.

Model 1 in (5.2.4) and Model 2 in (5.2.6) were used to measure the association between the log prothrombin index and the hazard rate. The algorithms described in Section 5.4 were used to estimate the parameters for Model 1 in (5.2.4) and for Model 2 in (5.2.6). We ran 100,000 iterations with a thinning of 4 and kept the last 5,000 iterations for making inferences.

First, the Gelman and Rubin convergence diagnostic were performed for the two models. A summary of the potential scale reduction factors for all the parameters are presented in Table 5.8. The results show that all the psrf point estimates for all the parameters are less than 1.2. Moreover, these point estimates are below the upper confidence limits. The results have confirmed that all chains have converged to the target posterior distributions. In addition, the Gelman and Rubin plots, the trace plots and ACF plots for all parameters in Model 1 and Model 2, presented in Appendix B, also confirm the convergences visually.

Table 5.8: Summary of MCMC convergence diagnostic tests for all of the parameters in Model 1.

Model 1			**Model 2**		
Potential scale reduction factors			Potential scale reduction factors		
	Point est.	Upper C.I.		Point est.	Upper C.I.
λ_1	1.01	1.06	λ_1	1.00	1.01
λ_2	1.03	1.15	λ_2	1.00	1.00
γ	1.01	1.03	γ	1.00	1.00
α	1.13	1.15	α	1.00	1.01
β_0	1.16	1.17	β_0	1.15	1.19
β_1	1.00	1.00	β_1	1.15	1.17
σ^2_ϵ	1.00	1.00	σ^2_ϵ	1.00	1.00
D_{11}	1.00	1.00	D_{11}	1.00	1.00
D_{12}	1.0	1.00	D_{22}	1.00	1.00
D_{21}	1.0	1.00	D_{33}	1.00	1.00
Multivariable psrf	1.15		Multivariable psrf	1.02	

A summary of results for the parameter estimates for Model 1 and Model 2 are presented in Table 5.9 and Table 5.10 respectively.

The fitted models are as follows:

Table 5.9: Summary statistics for parameter estimation of the liver cirrhosis data of Model 1 (5.2.4).

Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%
λ	4.236	0.237	3.726	4.115	4.246	4.373	4.705
γ	0.117	0.115	-0.113	0.040	0.115	0.195	0.345
α	-0.773	0.022	-0.817	-0.788	-0.773	-0.758	-0.726
β_0	4.207	0.006	4.196	4.203	4.208	4.211	4.220
β_1	0.017	0.002	0.013	0.016	0.017	0.018	0.022
σ	0.876	0.011	0.854	0.868	0.876	0.883	0.897
D_{11}	0.076	0.007	0.064	0.072	0.076	0.081	0.090
D_{12}	0.001	0.001	-0.001	0.0002	0.0006	0.0011	0.0019
D_{22}	0.002	0.001	0.0013	0.0015	0.0016	0.0017	0.0019

Table 5.10: Summary statistics for parameter estimation of the liver cirrhosis data of Model 2 (5.2.6).

Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%
λ_1	4.102	0.262	3.645	3.916	4.088	4.275	4.667
λ_2	0.011	0.009	0.001	0.003	0.007	0.015	0.034
γ	0.107	0.115	-0.123	0.029	0.107	0.187	0.329
α	-0.788	0.019	-0.825	-0.801	-0.788	-0.774	-0.751
β_0	4.281	0.006	4.268	4.276	4.281	4.285	4.293
β_1	-0.036	0.002	-0.041	-0.037	-0.036	-0.035	-0.032
σ	0.697	0.005	0.688	0.694	0.697	0.700	0.706
D_{11}	0.088	0.005	0.077	0.084	0.088	0.091	0.099
D_{22}	0.379	0.024	0.333	0.363	0.379	0.395	0.429
D_{33}	0.391	0.025	0.344	0.373	0.390	0.407	0.442

Table 5.11: The log-likelihood, AIC and BIC values for the fitted model 1 and fitted model 2.

	Log-likelihood	AIC	BIC
Model 1	-1690.771	3399.542	3437.255
Model 2	-1678.019	3378.038	3424.131

• Fitted model 1: The parameter estimates are taken from Table 5.9. The joint model with the hazard function at baseline having exponential distribution and linear mixed effects longitudinal submodel has the form

$$\begin{cases} \hat{h}_i(t) & = 4.236 \exp(0.117 * Treat - 0.773 * \hat{m}_i(t)) \\ \hat{m}_i(t) & = 4.207 + 0.017 * t + \hat{u}_{i0} + \hat{u}_{i1} * t \,. \end{cases}$$

• Fitted model 2: The parameter estimates are taken from Table 5.10. The joint model with the hazard function at baseline having Gompertz distribution and nonlinear mixed effects longitudinal submodel has the form

$$\begin{cases} \hat{h}_i(t) & = 4.102 * \exp(0.011 * t) \exp(0.107 * Treat - 0.788 * \hat{m}_i(t)) \\ \hat{m}_i(t) & = 4.281 - 0.036 * t + \hat{u}_{i1} * (t - \mathcal{K}_1) + \hat{u}_{i2} * (t - \mathcal{K}_2)6 + \hat{u}_{i3} * (t - \mathcal{K}_3) \,, \end{cases}$$

where the three fitted knots are placed at 25%, 50% and 75% of the observed times.

The maximized log-likelihood, AIC and BIC values of the two fitted models are presented in Table 5.11. The results show that the fitted model 2 improved the log-likelihood when we put knots in the longitudinal submodel. In a similar way, both the AIC and BIC values of the fitted model 2 are lower than the fitted model 1. These results confirm that the penalized spline joint model is the better fitted model.

5.8 Discussion

In this chapter, a fully Bayesian analysis was implemented for the penalized spline joint models. We also proposed an algorithm which combines the GS, random walk MH and

independent MH algorithms. A set of R code following the main algorithm was written for the two joint models. In the first joint model, the random effects part has two dimensions with a non-diagonal covariance matrix. In the second joint model, the random effects part has four dimensions with a diagonal covariance matrix. To validate the proposed algorithm, two simulation studies were implemented. The results show that: (i) the computational task for this approach is feasible and simple because the algorithms avoid multi-integral calculation; (ii) the uncertainties are inferred fully for each parameter in the models.

Another important aim of this chapter was to assess the sensitivity to prior specification of the baseline hazard rate parameter and association parameter. The results from the prior sensitivity analysis show that the estimates are insensitive to different prior distributions of these parameters. In addition, the coverage performances were extremely good for the survival parameters under different prior specifications.

The drawbacks observed through the simulation studies were: (i) the computing time for the Bayesian analysis was not yet efficient in time. This might be due to the fact that the code was run on a normal desk computer with an Intel (R), Core(TM) i7-3770 CPU (3.40 GHz) and 8 GB RAM running Windows XP. We also had to run many MCMC iterations for a long time to ensure the convergence to the target posterior distribution; (ii) when the dimension of random effects was high, the independent MH algorithm led to a larger variation in the parameter samples.

Chapter 6

Summary and Future Direction

6.1 Achieved aims

The aims mentioned in Chapter 1 have been achieved through the exposition in Chapters 3, 4, and 5. In Chapter 3, the penalized spline joint models were introduced to handle non-linear longitudinal and survival data. We also proposed a full likelihood approach for estimating parameters in these joint models. Simulation studies showed that these models can flexibly fit and well predict the association of the two types of data. The original contributions in Chapter 3 are as follows:

(i) The two groups of penalized spline joint models were presented namely the penalized spline joint models with hazard rate at baseline having a Gompertz distribution and the penalized spline joint models with a piecewise-constant baseline risk function;

(ii) The parameter estimation for the penalized spline joint models was proposed. This estimation method is viewed as the full likelihood approach;

(iii) The ECM algorithm is presented;

(iv) R code following the ECM algorithm was written for the two groups of penalized spline joint models.

In Chapter 4, the theory for a new two-stage approach was introduced for penalized spline joint models. The method can improve the computational problem of the full likelihood method in Chapter 3 by approximating the log-likelihood function. Parameter estimation was quick and effective compared to the ordinary two-stage approach and the full likelihood approach in simulation studies. The following are the achievements from this chapter:

(i) A modified two-stage approach and an approximation theorem were proposed for the penalized spline joint models;

(ii) The estimation algorithm was detailed in two stages for longitudinal parameters and survival parameters;

(iii) The random effects misspecification analysis was investiagted;

(iv) R code was written and used for extensive simulation studies and a case study.

In Chapter 5, we applied a fully Bayesian approach to the penalized spline joint models. This method avoided the multi-integral calculation from the full likelihood method and the approximation from the two-stage method. Moreover, a sample of target posterior distribution for each parameter is inferred using a combination of well known algorithms. It can also open the way to reduce the impact of normality assumption for random effects from Chapter 3 and Chapter 4. Specific contributions were:

(i) The joint prior distributions and joint posterior distributions for the penalized spline joint models were proposed;

(ii) The main algorithm was presented with detailed sub-algorithms;

(iii) Prior sensitivity analysis was conducted for the parameter of baseline hazard rate and the association parameter;

(iv) R code was written to check for the validity of the algorithm. Moreover, the code was used for a case study.

6.2 Limitations

In summary, three approaches were proposed to estimates parameters for penalized spline joint models namely, a full likelihood approach, a modified two-stage approach and a fully Bayesian approach. In each approach, there are advantages and disadvantages for estimating parameters in the penalized spline joint models. Through simulation studies from the three main chapters, we observed some limitations for the penalized spline joint models and for the proposed parameter estimation methods.

The representation of the polynomial basis in the penalized spline joint models can give an intuitive and easy way to model non-linear longitudinal covariates. However, there are some drawbacks with the polynomial basis. For very nonuniform knots, the truncated power functions may form an ill-conditioned basis De Boor (1978) and Dierckx (1995). In this case, some of the basis functions become nearly linearly dependent on the others. This leads to unsuitable numerical calculations. Moreover, in order to model for a non-linear individual curve, a large number of knots need to be inserted into the longitudinal submodel. The joint modelling becomes complicated very quickly with high dimensions of random effects.

The maximisation of the observed data log-likelihood functions for the penalized spline joint models is often intractable. Therefore, algorithms or special techniques need to be applied to approximate the solution. In the full likelihood approach in Chapter 3, there are no closed-form solutions to the integrals with respect to the random effects as well as for the integrals with respect to time in the survival function under the ECM algorithm. This leads to complicated computational problems. As a result, the algorithm is very time consuming when dealing with high dimensions of random effects. Another drawback for both the full likelihood method and the two-stage method is that the estimation results depend too much on the normality assumptions of random effects and error terms. In addition, the random effects and longitudinal parameters in the proposed two-stage approach are estimated from the separate linear mixed model. This can cause biases when data have high imformative dropout.

The fully Bayesian approach is also time consuming when flat prior distributions are applied for parameters. The independent MH algorithm is used to update random effects in the penalized spline joint models. However, this algorithm has limitations when the dimension of random effects is large. In addition, the conditional posterior distribution for random effects becomes complex in this case. These lead to inaccurate estimation for the random effects and the standard deviation of the estimates is very large.

6.3 Future direction

Based on the above limitations, our future work will focus on more flexible penalized joint models with a different basis to handle non-linear longitudinal data. The B-spline

basis is one that can remove the problems associated with the polynomial basis. However, the application of penalized B-splines for joint models will be more complex in terms of modelling and estimating. In addition, missing data mechanisms need to be considered for joint models in order to obtain accurate results. Different types of censoring mechanisms need to apply different inferential procedures.

In the classical approach, the future work is on reducing computational complexity. New approximation methods need to be applied for multi-integrals to minimize the errors and convergence time. In addition, we will study the effects of random effects misspecification on these new methods and try to relax the assumptions for the random effects distribution and the error measurement distribution.

To improve the weaknesses in the modified two-stage approach, we will combine this approach with regression calibration approaches. By doing this, the informative dropout will be acounted for estimating the parameters in the longitudinal model. This approach will be considered as an extension for the works of Ye et al. (2008) .

Based on the limitations of the Bayesian approach, new algorithms need to be proposed for flexible penalized joint models to improve convergence time. We will focus on relaxing the prior information of the random effects in our future work. Furthermore, we will replace the independent MH algorithm to estimate the random effects precision matrix with an effort to improve the biases and variation. The prior sensitivity analysis needs to be extended for all of parameters in the joint models.

Bibliography

O. Aalen, O. Borgan, and H. K. Gjessing. *Survival and Event History*. Springer, 2008.

D. I. Abrams, A. I. Goldman, C. Launer, J. A. Korvick, J. D. Neaton, L. R. Crane, M. Grodesky, S. Wakefield, K. Muth, and S. Kornegay. Comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus infection who are intolerant of or have failed zidovudine therapy. *New England Journal of Medicine*, 330: 657–662, 1994.

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. SpriSpringer-Ve, New York, 1993.

P. C. Austin. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29):3946–3958, 2012.

D. Bates, M. maechler, and B. Bolker. Linear mixed-effects models using S4 classes. r package version 0.999375-42. *http://cran.r-project.org/pakage=lme4*, 2011.

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

S. Brooks, A. Gelman, G. L. Jones, and X. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

E. Brown and J. Ibrahim. A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, 59(2):221–228, 2003.

E. R. Brown, J. G. Ibrahim, and V. DeGruttola. A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73, 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.

P. Bycott and J. Taylor. A comparison of smoothing techniques for CD4 data measured with error in a time dependent Cox proportional hazards model. *Statistics in Medicine*, 17(18):2061–2077, 1998.

D. Cox and D. Oakes. *Analysis of Survival Data*. Chapman and Hall, London, 1984.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall/CRC Press, New York, 1979.

M. J. Crowther and P. C. Lambert. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32(23):4118–4134, 2013.

I. D. Currie and M. Durban. Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, 2(4):333–349, 2002.

U. G. Dafni and A. A. Tsiatis. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54(4):1445–1462, 1998.

C. De Boor. *A Practical Guide to Splines*, volume 27. Springer-Verlag New York, 1978.

P. Dierckx. *Curve and Surface Fitting with Splines*. Oxford University Press, 1995.

P. Diggle, P. Heagerty, K. Y. Liang, and S. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, New York, 2002.

J. Ding and J. Wang. Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, 64(2):546–556, 2008.

M. Durban, J. Harezlak, MP. Wand, and RJ. Carroll. Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24(8):1153–1167, 2005.

P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

C. L. Faucett and D. C. Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15 (15):1663–1685, 1996.

G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied Longitudinal Analysis*, volume 998. John Wiley & Sons, 2004.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/ Hierarchical Models*, volume 1. Cambridge University Press, New York, 2007.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London, 1995.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

I. J. Good and R. A. Gaskins. Noparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

A. Gould, M. E. Boye, M. J. Crowther, J. G. Ibrahim, G. Quartey, S. Micallef, and F. Y. Bois. Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the DIA Bayesian joint modeling working group. *Statistics in Medicine*, 34 (14):2181–2195, 2014.

D. J. Hand and M. J. Crowder. *Practical Longitudinal Data Analysis*, volume 34. CRC Press, 1996.

F. Harrell. *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, 2015.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

Philip Hougaard. *Analysis of Multivariate Survival Data*. Springer-Verlag, New York, 2000.

X. Huang. Diagnosis of random effect model misspecification in generalized linear mixed models for binary response. *Biometrics*, 65(2):361–368, 2009.

P. Huong, D. Nur, and A. Branford. Penalized spline joint models for longitudinal and time-to-event data. *Communication in Statistics Theory and Methods*, 2016.

J. G. Ibrahim, M. Chen, and D. Sinha. *Bayesian Survival Analysis*. Wiley Online Library, 2005.

J. Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer-Verlag, New York, 2010.

J. Kalbfleisch and R. Prentice. *The Statistical Analysis of Failure Time Data, 2nd edition*. Wiley, New York, 2002.

J. P. Klein and M. L. Moeschberger. *Survival Analysis - Techniques For Censored and Truncated Data*. Springer-Verlag, New York, 2005.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

M. J. Lindstrom and D. M. Bates. Newton-raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982.

C. McCulloch, S. Searle, and J. Neuhaus. *Generalized, Linear and Mixed Models*. Wiley, New Jersey, 2008.

C. E. McCulloch and J. M. Neuhaus. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science*, pages 388–402, 2011.

G. McLachlan and TH. Krishnan. *The EM Algorithm and Extensions*, volume 382. John Wiley & Sons, 2007.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6):1087–1092, 1953.

G. Molenberghs and G. Verbeke. *Models for Discrete Longitudinal Data*. Springer-Verlag, New York, 2005.

P. A. Murtaugh, E. R. Dickson, G. M. Van Dam, M. Malinchoc, P. M. Grambsch, A. L. Langworthy, and C. H. Gips. Primary biliary cirrhosis: Prediction of short term survival based on repeated patient visits. *Hepatology*, 20(1):126–134, 1994.

J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar. R core team (2014) nlme: linear and nonlinear mixed effects models. r package version 3.1-117. *Available at h ttp://CRAN. R-project. org/package= nlme*, pages –, 2014.

P. Rao. *Variance Components: Mixed Models, Methodologies and Applications*, volume 78. Chapman & Hall/CRC Press, Boca Raton, 1997.

J. A. Rice and C. O. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, pages 253–259, 2001.

D. Rizopoulos. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35 (9):1–33, 2010.

D. Rizopoulos. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule. *Computational Statistics and Data Analysis*, 56:491–501, 2011.

D. Rizopoulos. *Joint Models for Longitudinal And Time-to-event Data with Applications in R*. Chapman & Hall/CRC, Biostatistics series, 2012.

D. Rizopoulos. The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *arXiv preprint arXiv:1404.7625*, pages –, 2014.

C. P. Robert and G. Casella. Monte carlo statistical methods. *New York*, pages –, 2004.

D. Ruppert, M. Wand, and R. Carroll. *Semiparametric Regression*. Cambridge University Press, Cambridge, 2003.

D. Ruppert, M. P. Wand, and R. J. Carroll. Semiparametric regression during 2003 to 2007. *Electronic Journal of Statistics*, 3:1193–1256, 2009.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

S. Self and Y. Pawitan. Modeling a marker of disease progression and onset of disease. pages 231–255, 1992. ISSN 1475712316.

J. D. Singer and J. B. Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University press, 2003.

M. J. Sweeting and S. G. Thompson. Joint modelling of longitudinal and time to event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5):750–763, 2011.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending The Cox Model*. Springer-Verlag, New York, 2000.

T. M. Therneau, T. M.Therneau. The survival package for r. *Accessed at http://CRAN.R-project.org/package=survival*, pages –, 2014.

A. A. Tsiatis and M. Davidian. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, pages 447–458, 2001.

A. A. Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004. ISSN 1017-0405.

A. A. Tsiatis, V. Degruttola, and M. S. Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37, 1995.

W. N. Venables and B. D. Ripley. Modern applied statistics with S-plus. pages –, 2013.

G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York, 2000.

S. Viviani, M. Alfo, and D. Rizopoulos. Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing*, 24(3):417–427, 2014.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Jon Wakefield. *Bayesian and Frequentist Regression Methods*. Springer Science & Business Media, 2013.

Y. Wang and J. M. G. Taylor. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of The American Statistical Association*, 96(455):895–905, 2001.

W. Ye, X. Lin, and J. MG. Taylor. Semiparametric modeling of longitudinal measurements and time- to- event data: a two stage regression calibration approach. *Biometrics*, 64 (4):1238–1246, 2008.

Appendices

A. Appendices for Chapter 3

A.1 Simulated data of the penalized spline joint model

One sample of simulated data of the penalized spline joint model in (3.4.1) is presented in Table A.1 for the first three patients. The subjects were measured bimonthly and the entry time was 0 for all subjects. The *Obstime* variable includes the time points at which these measurements were recorded. The *Time* variable includes the observed survival times when the subject meets an event. x is a time-constant binary random variable with parameter $p = 0.5$. Column y contains the longitudinal responses. The *Death* variable is the event status indicator. This variable receives value 1 when the true survival time is less than or equal to the censoring time and 0 otherwise. We define the four random effects variables which are $Z_1 = (obstime - \mathcal{K}_1)_+$, $Z_2 = (obstime - \mathcal{K}_2)_+$, $Z_3 = (obstime - \mathcal{K}_3)_+$, and $Z_4 = \mathbf{1}$. For the longitudinal process, there are 1902 observations for 500 subjects. For each subject, 1-7 longitudinal measurements are recorded. On average, there are four longitudinal measurements per subject. For the event process, there are 297 subjects who meet for an event which is equivalent to 59.4 % of the whole sample.

Table A.1: A snapshot of simulated data for penalized spline joint model in (3.4.1).

Id	Obstime	Time	x	y	Death	Z_1	Z_2	Z_3	Z_4
1	0.0	4.97	0	1.41	1	0.0	0.0	0.0	1
1	0.5	4.97	0	6.45	1	0.0	0.0	0.0	1
1	1.0	4.97	0	4.10	1	0.0	0.0	0.0	1
1	1.5	4.97	0	1.50	1	0.5	0.0	0.0	1
1	2.0	4.97	0	4.07	1	1.0	0.0	0.0	1
1	2.5	4.97	0	6.16	1	1.5	0.5	0.0	1
1	3.0	4.97	0	3.60	1	2.0	1.0	0.0	1
1	3.5	4.97	0	8.32	1	2.5	1.5	0.5	1
1	4.0	4.97	0	6.32	1	3.0	2.0	1.0	1
2	0.0	2.79	0	6.81	1	0.0	0.0	0.0	1
2	0.5	2.79	0	7.77	1	0.0	0.0	0.0	1
2	1.0	2.79	0	9.75	1	0.0	0.0	0.0	1
2	1.5	2.79	0	11.04	1	0.5	0.0	0.0	1
2	2.0	2.79	0	7.20	1	1.0	0.0	0.0	1
3	0.0	1.90	0	-1.84	0	0.0	0.0	0.0	1
3	0.5	1.90	0	1.12	0	0.0	0.0	0.0	1
3	1.0	1.90	0	0.78	0	0.0	0.0	0.0	1

A.2 The updating rule for the parameters

The integrals with respect to the random effects in (3.3.7) do not have closed-form solutions. Therefore, in this paper, we implement the Gaussian-Hermite quadrature rule as in Rizopoulos (2011) to approximate the integrals. In our simulation study and R coding, we use the Gaussian-Hermite quadrature rule with 10 quadrature points. The updating formulas of the parameters in Step 3 have different forms for each parameter following Rizopoulos (2012). We have the closed-form estimates for the measurement error variance σ_ε^2 in the longitudinal model and the covariance matrix of the random effects as follows

$$\hat{\boldsymbol{G}}^{(it+1)} = \frac{1}{n}\sum_i \int \boldsymbol{b}_i^T \boldsymbol{b}_i p(\boldsymbol{b}_i \mid T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i = \frac{1}{N}\sum_i v\tilde{\boldsymbol{b}_i^{(it)}} + \tilde{\boldsymbol{b}_i^{(it)}}\tilde{\boldsymbol{b}_i^{(it)}}^T, \qquad (.1)$$

where $\tilde{\boldsymbol{b}_i} = E(\boldsymbol{b}_i|T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}) = \int \boldsymbol{b}_i p(\boldsymbol{b_i}|T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}) d\boldsymbol{b}_i$ and $v\tilde{\boldsymbol{b}_i} = var(\boldsymbol{b}_i|T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}) = \int (\boldsymbol{b}_i - \tilde{\boldsymbol{b}_i}) p(\boldsymbol{b_i}|T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}) d\boldsymbol{b}_i$. The updating formula for σ_ε^2 is

$$\hat{\sigma_\varepsilon^2}^{(it+1)} = \frac{1}{n}\sum_i \int \boldsymbol{W}^T \boldsymbol{W} p(\boldsymbol{b}_i \mid T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i \quad , \qquad (.2)$$

where $\boldsymbol{W} = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{X}_i\boldsymbol{u}_i - \boldsymbol{Z}_i\boldsymbol{v}_i$. Unfortunately, we cannot obtain closed-form expressions for the fixed effects $\boldsymbol{\beta}$ and the parameters of the survival submodel γ, α, *and* $\boldsymbol{\theta}_{h_0}$. Thus we employ the one-step Newton-Raphson approach to obtain the updated $\boldsymbol{\beta}^{(it+1)}$, $\gamma^{(it+1)}$,$\alpha^{(it+1)}$ and $\boldsymbol{\theta}_{h_0}^{(it+1)}$. In particular,

$$\begin{aligned} S(\boldsymbol{\theta}) &= \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(it)})}{\partial \boldsymbol{\theta}} \\ \hat{\boldsymbol{\theta}}^{(it+1)} &= \hat{\boldsymbol{\theta}}^{(it)} - \left[\frac{\partial S(\hat{\boldsymbol{\theta}}^{(it)})}{\partial \boldsymbol{\theta}}\right]^{-1} S(\hat{\boldsymbol{\theta}}^{(it)})\,, \end{aligned} \tag{.3}$$

where $S(\boldsymbol{\theta})$ is the score vector corresponding to parameter $\boldsymbol{\theta}$ and the score vector has the form

$$\begin{aligned} S(\boldsymbol{\theta}) &= \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(it)})}{\partial \boldsymbol{\theta}} \\ &= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}^T} \log\left\{p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}^{(it)}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}^{(it)}) p(\boldsymbol{b}_i; \boldsymbol{\theta}^{(it)})\right\} . p(\boldsymbol{b}_i \mid T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i\,. \end{aligned}$$

A.3 Simulating survival time

There are four cases for simulating survival time T_i of the model (3.4.1) as follows.

When the survival time $t < \mathcal{K}_1$, we calculate the cummulative hazard function $H_i(t) = \int_0^t h_i(s)ds$. Based on the relation between the survival function $S_i(t)$, cummulative hazard function $H_i(t)$ and cummulative distribution $F_i(t)$, we have

$$S_i(t) = \exp(-H_i(t)) = 1 - F_i(t)\,.$$

Following this result, we set

$$u = 1 - F_i(T_i)\,,$$

where u is a random variable with $u \sim \mathcal{U}(0,1)$. The survival time t is the solution of the equation

$$U = \exp(-H_i(t)) = \exp(-\int_0^t h_i(s)ds)\,.$$

The condition $t < \mathcal{K}_1$ is equal to

$$-\log(U) < \int_0^{\mathcal{K}_1} h(s)ds\,.$$

When $\mathcal{K}_1 \leq t < \mathcal{K}_2$, we calculate the cummulative hazard function $H_i(t) = \int_0^{\mathcal{K}_1} h_i(s)ds + \int_{\mathcal{K}_1}^{t} h_i(s)ds$. The survival time t is the solution of the equation

$$U = \exp\left[-\left\{\int_0^{\mathcal{K}_1} h_i(s)ds + \int_{\mathcal{K}_1}^{t} h_i(s)ds\right\}\right],$$

where U is a value of $u \sim \mathcal{U}(0,1)$. The condition $\mathcal{K}_1 \leq t < \mathcal{K}_2$ is equal to

$$-\log(U) < \int_0^{\mathcal{K}_1} h_i(s)ds + \int_{\mathcal{K}_1}^{\mathcal{K}_2} h_i(s)ds\,.$$

When $\mathcal{K}_2 \leq t < \mathcal{K}_3$, we calculate the cummulative hazard function $H_i(t) = \int_0^{\mathcal{K}_1} h_i(s)ds + \int_{\mathcal{K}_1}^{\mathcal{K}_2} h_i(s)ds + \int_{\mathcal{K}_2}^{t} h_i(s)ds$. The survival time t is the solution of the equation

$$U = \exp\left[-\left\{\int_0^{\mathcal{K}_1} h_i(s)ds + \int_{\mathcal{K}_1}^{\mathcal{K}_2} h_i(s)ds + \int_{\mathcal{K}_2}^{t} h_i(s)ds\right\}\right],$$

where U is a value of $u \sim \mathcal{U}(0,1)$. The condition $\mathcal{K}_2 \leq t < \mathcal{K}_3$ is equal to

$$-\log(U) < \int_0^{\mathcal{K}_1} h_i(s)ds + \int_{\mathcal{K}_1}^{\mathcal{K}_2} h_i(s)ds + \int_{\mathcal{K}_2}^{\mathcal{K}_3} h_i(s)ds\,.$$

When $\mathcal{K}_3 \leq t$, the cummulative hazard function has the form $H_i(t) = \int_0^{\mathcal{K}_1} h_i(s)ds + \int_{\mathcal{K}_1}^{\mathcal{K}_2} h_i(s)ds + \int_{\mathcal{K}_2}^{\mathcal{K}_3} h_i(s)ds + \int_{\mathcal{K}_3}^{t} h_i(s)ds$. The survival time t is the solution of the equation

$$U = \exp\left[-\left\{\int_0^{\mathcal{K}_1} h_i(s)ds + \int_{\mathcal{K}_1}^{\mathcal{K}_2} h_i(s)ds + \int_{\mathcal{K}_2}^{\mathcal{K}_3} h_i(s)ds + \int_{\mathcal{K}_3}^{t} h_i(s)ds\right\}\right].$$

A.4 Summary statistics for parameter estimation

Table A.2: Summary statistics for parameter estimation of the simulated data of the model in (3.4.4) for different censoring rates.

Parameter	True value	Censored (20%)			Censored (40%)		
		Estimate	SD	MSE	Estimate	SD	MSE
β_0	5	4.85	0.30	0.25	5.10	0.30	0.27
β_1	2	1.86	0.45	0.20	2.10	0.57	0.18
λ_1	0.1	0.13	0.12	0.00	0.11	0.10	0.00
λ_2	0.5	0.52	0.07	0.00	0.49	0.14	0.02
γ	0.5	0.48	0.10	0.00	0.51	0.09	0.00
α	0.05	0.05	0.02	0.00	0.04	0.04	0.00
σ	2	2.02	0.05	0.00	2.02	0.06	0.00
D_{11}	2	2.21	0.67	0.17	2.27	0.80	0.22
D_{22}	2	2.16	0.27	0.09	2.10	0.43	0.05
D_{33}	2	2.26	0.27	0.01	2.22	0.60	0.10
D_{44}	4	4.20	0.53	0.20	4.24	0.63	0.18

B. Plots for case study in Chapter 5.

B.1 Gelman and Rubin diagnostic plots in Model 1.

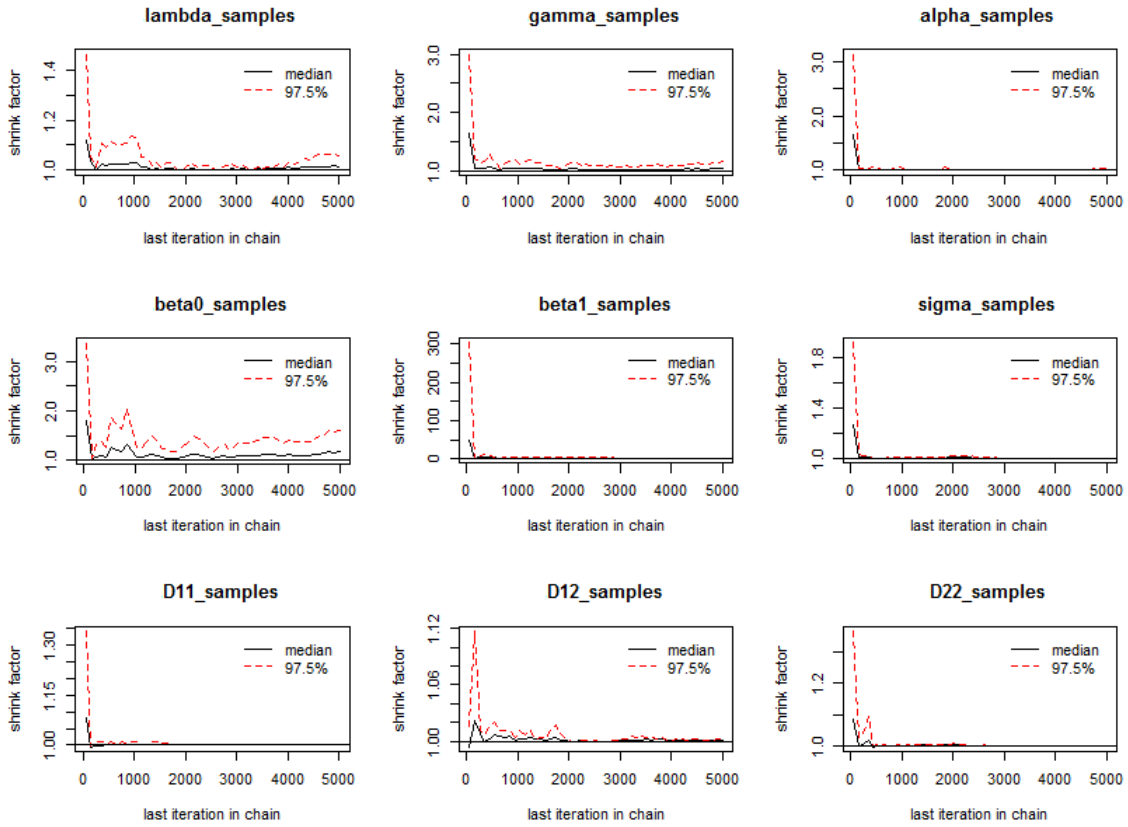

Figure B1.1: The potential rate reduction factor plots of Gelman and Rubin diagnostic for all the parameters in Model 1.

B.2 Gelman and Rubin diagnostic plots in Model 2.

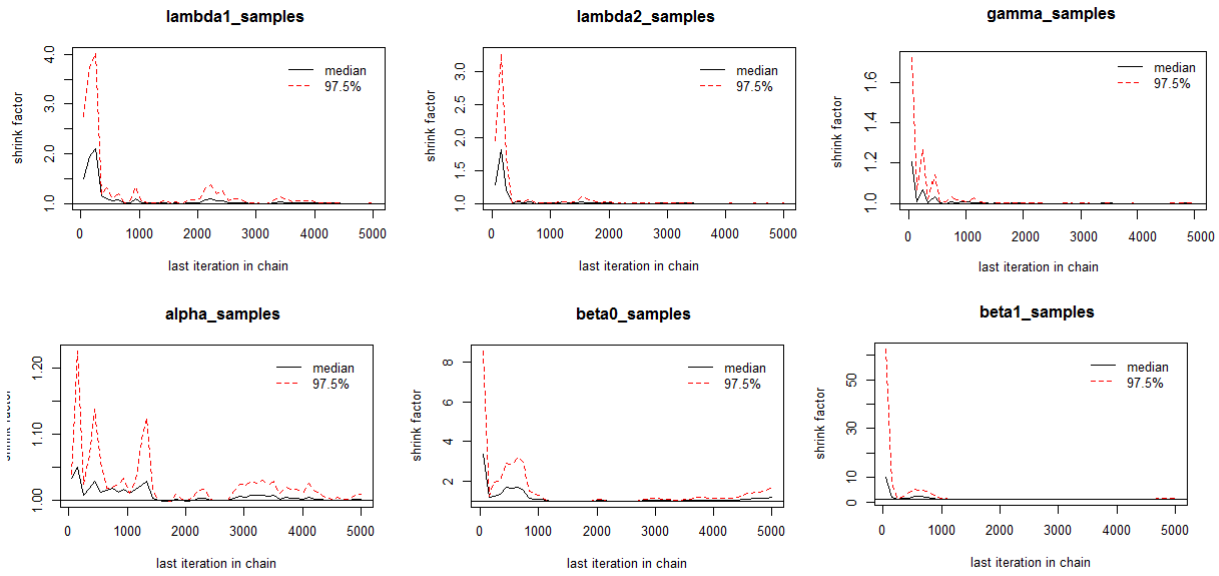

Figure B2.1: The potential rate reduction factor plots of Gelman and Rubin diagnostic for the parameters λ_1, λ_2, γ, α, β_0 and β_1 in Model 2.

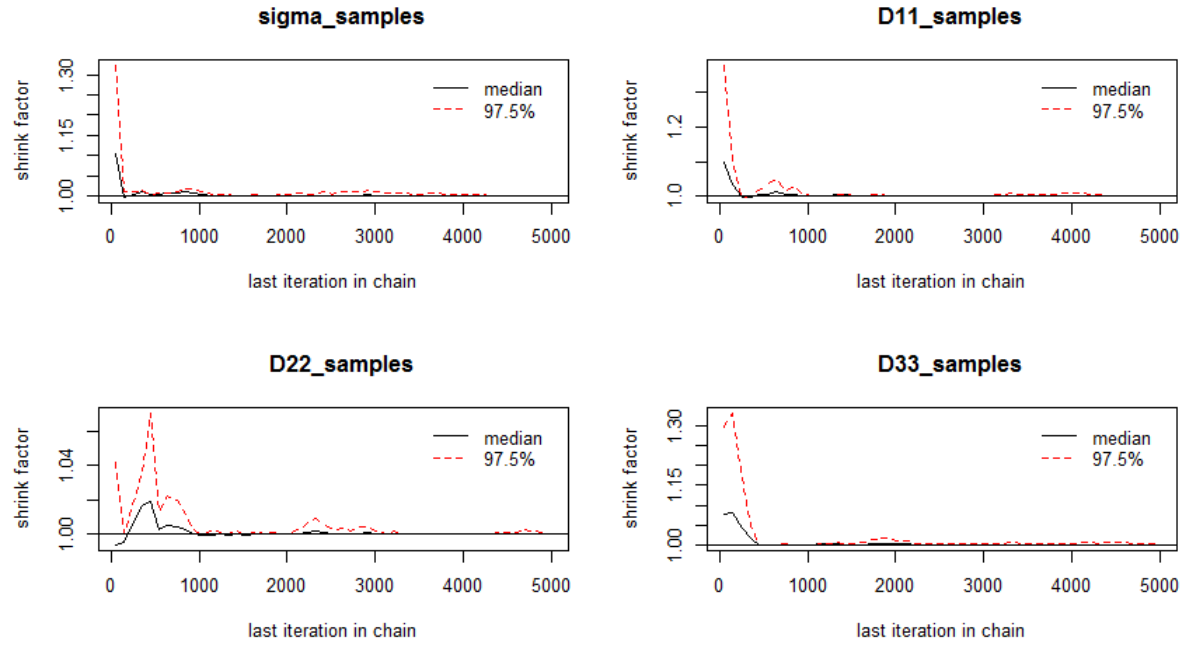

Figure B2.2: The potential rate reduction factor plots of Gelman and Rubin diagnostic for the parameters σ_ε^2, D_{11}, D_{22} and D_{33} in Model 2.

B.3 ACF plots in Model 1 and Model 2.

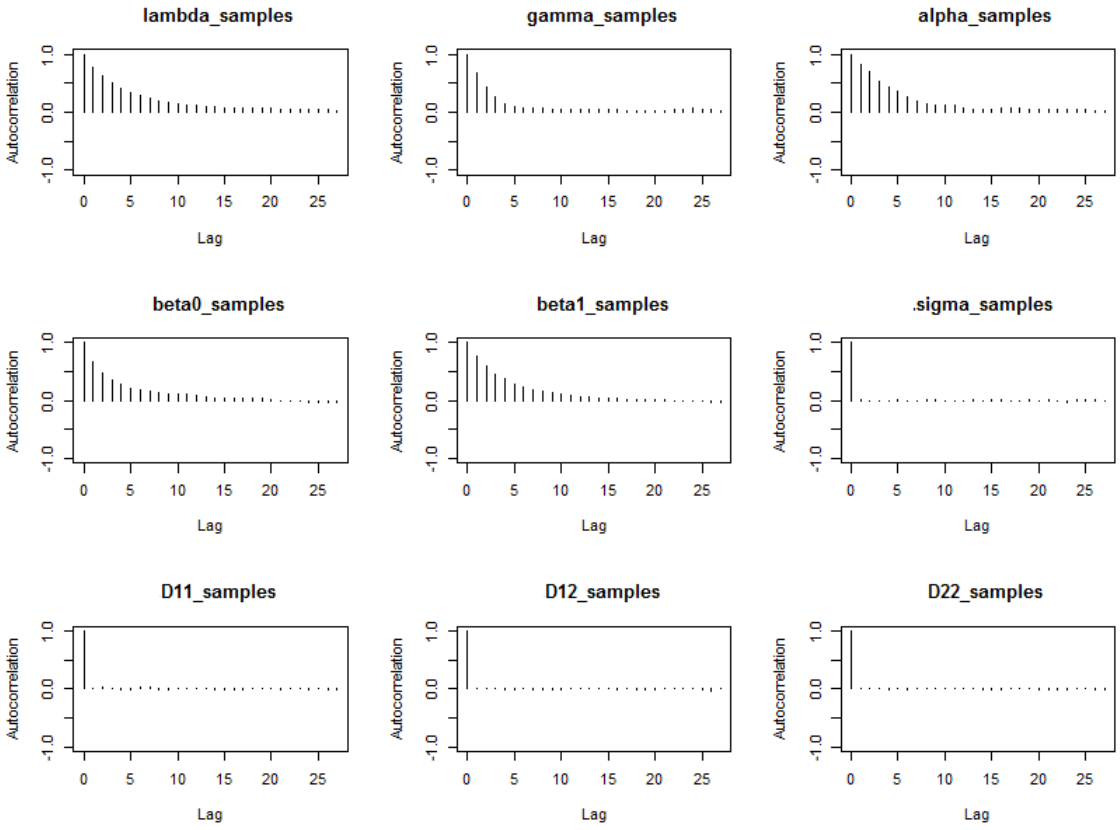

Figure B3.1: ACF plots for all the parameters in Model 1.

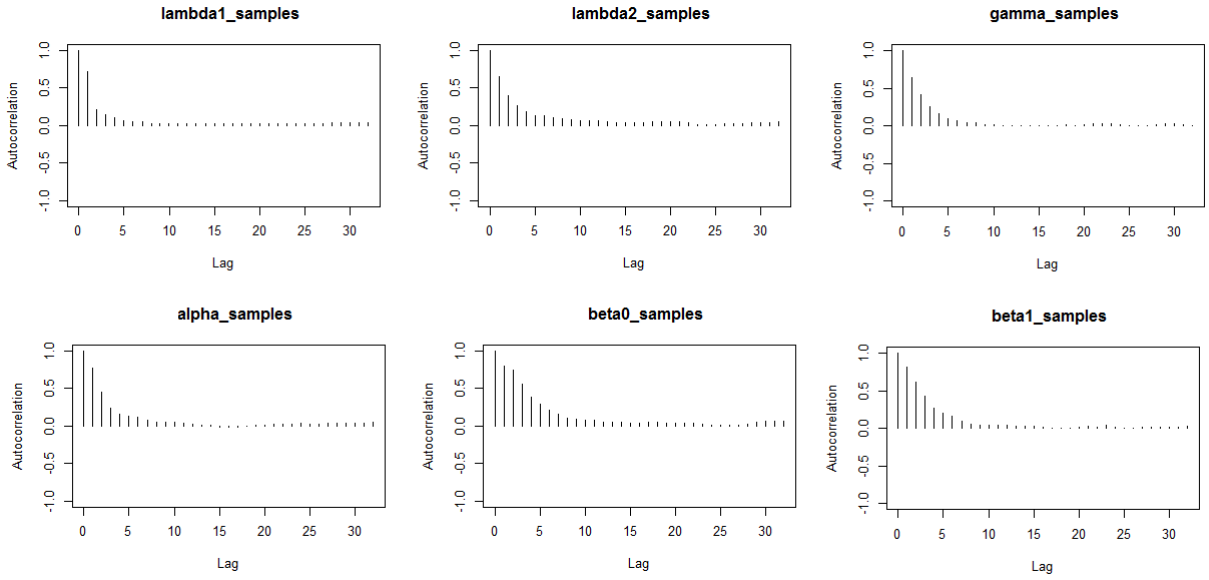

Figure B3.2: ACF plots for the parameters λ_1, λ_2, γ, α, β_0 and β_1 in Model 2.

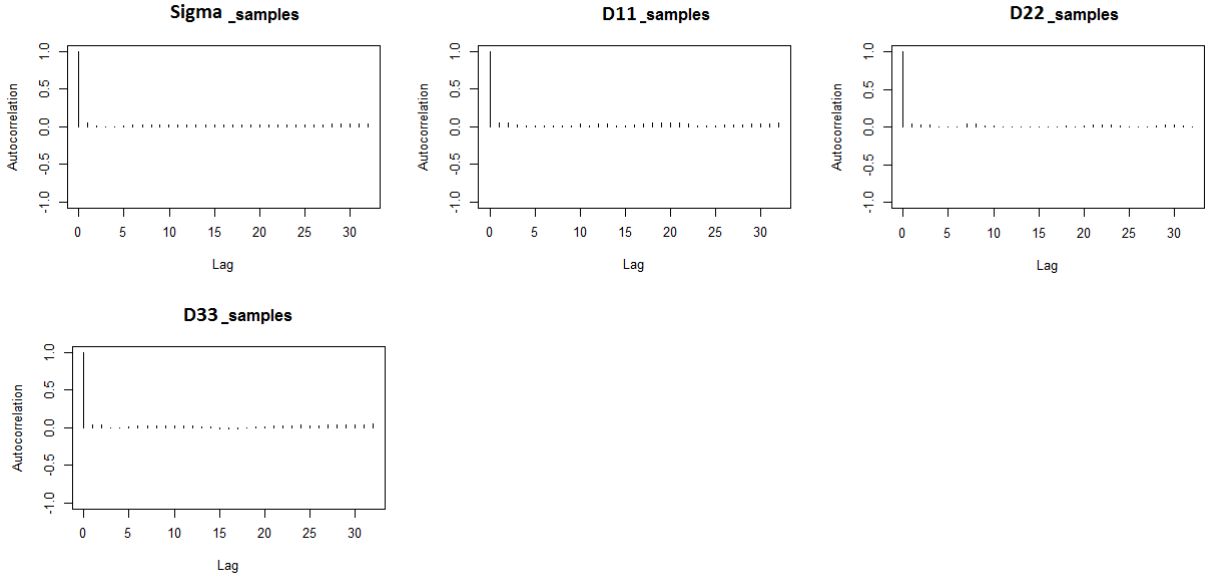

Figure B3.3: ACF plots for the parameters σ_ε^2, D_{11}, D_{22} and D_{33} in Model 2.

B.4 MCMC traces and posterior distribution plots in Model 1 and Model 2.

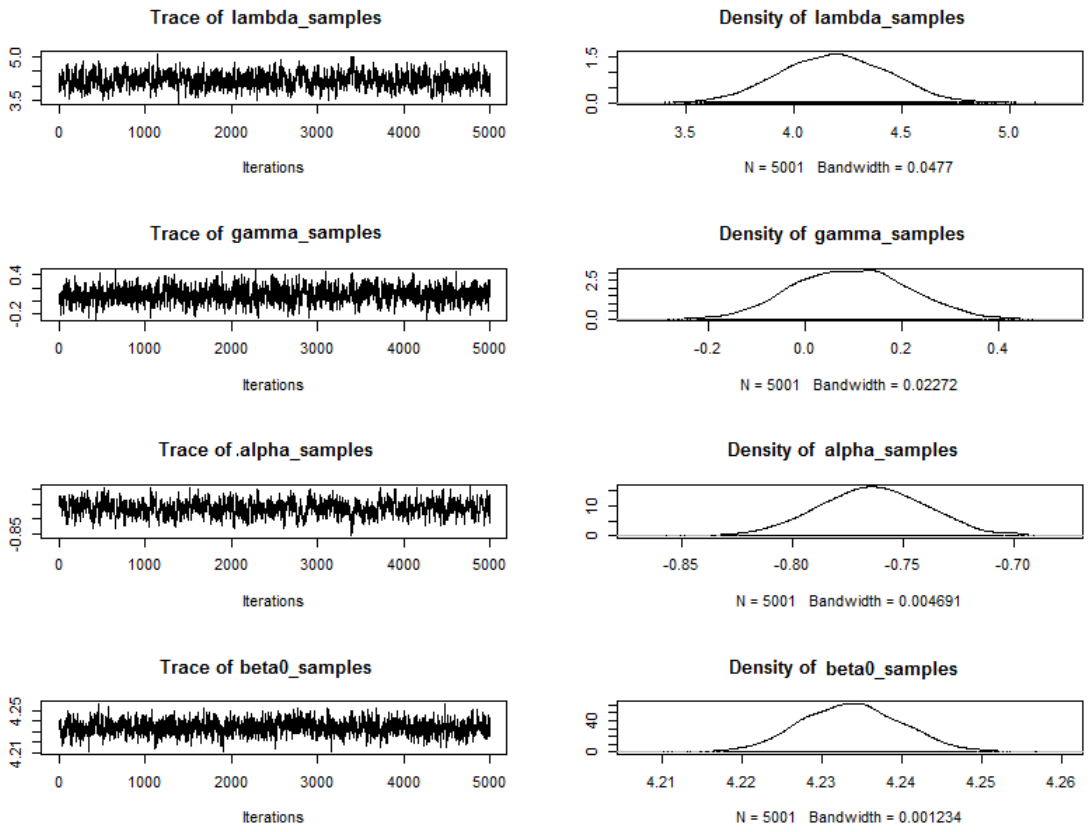

Figure B4.1: MCMC traces and posterior distribution plots for the parameters λ, γ, α and β_0 in Model 1.

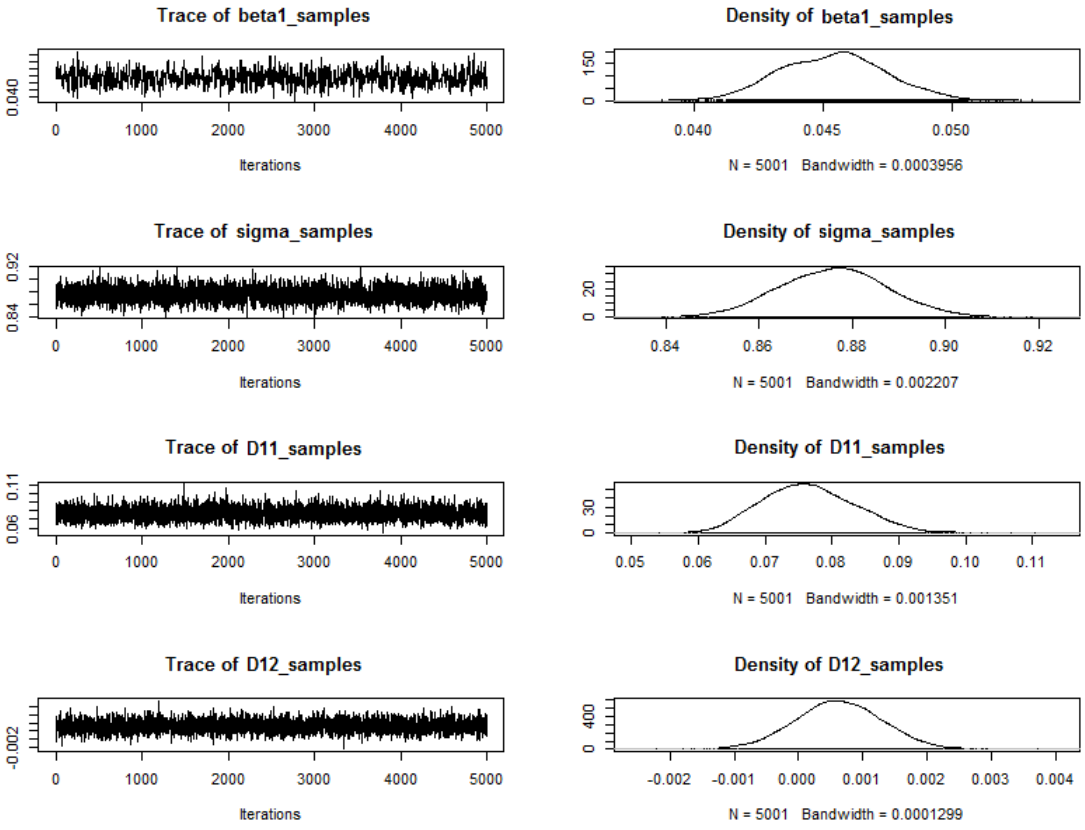

Figure B4.2: MCMC traces and posterior distribution plots for the parameters β_1, σ_ε^2, D_{11} and D_{212} in Model 1.

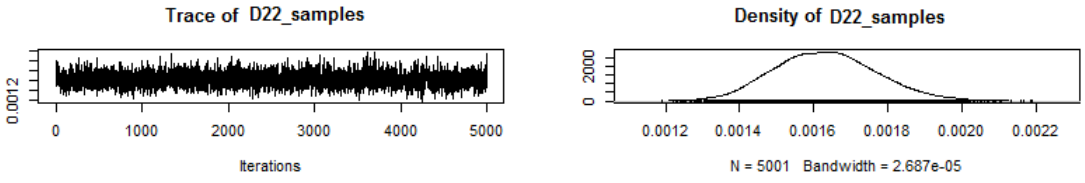

Figure B4.3: MCMC traces and posterior distribution plots for the parameter D_{22} in Model 1.

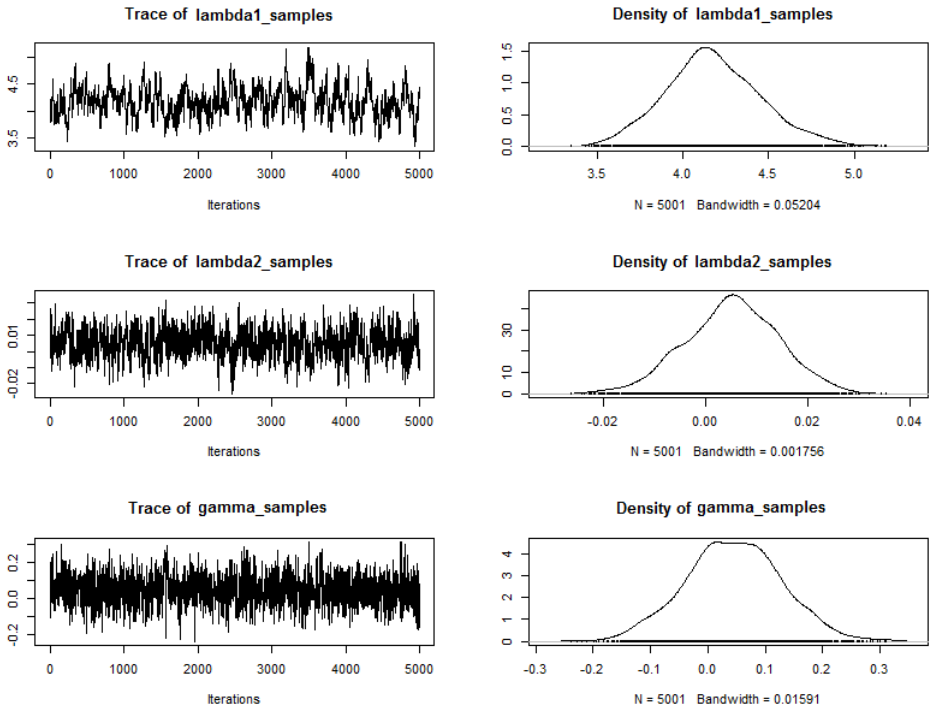

Figure B4.4: MCMC traces and posterior distribution plots for the parameters λ_1, λ_2 and γ in Model 2.

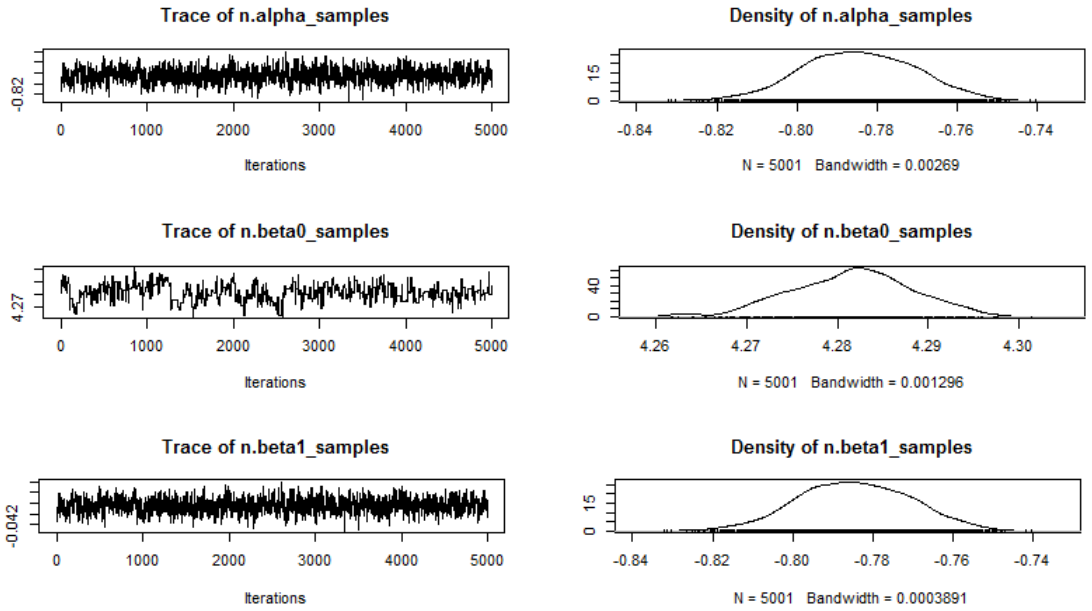

Figure B4.5: MCMC traces and posterior distribution plots for the parameters α, β_0 and β_1 in Model 2.

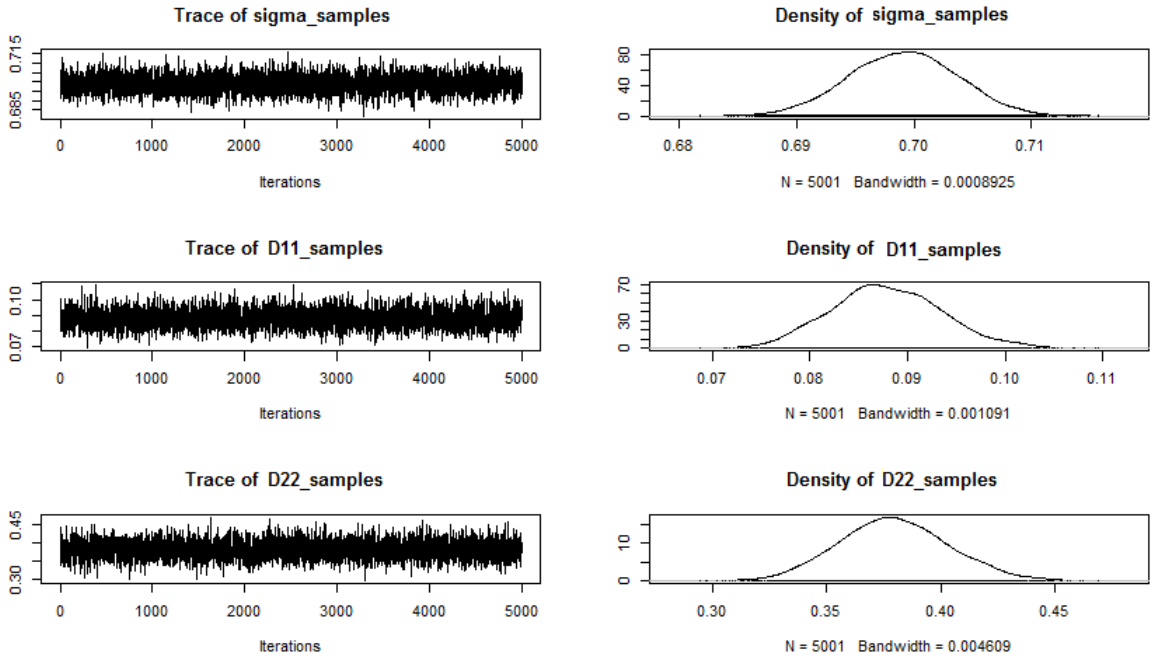

Figure B4.6: MCMC traces and posterior distribution plots for the parameters σ_ε^2, D_{11} and D_{22} in Model 2.

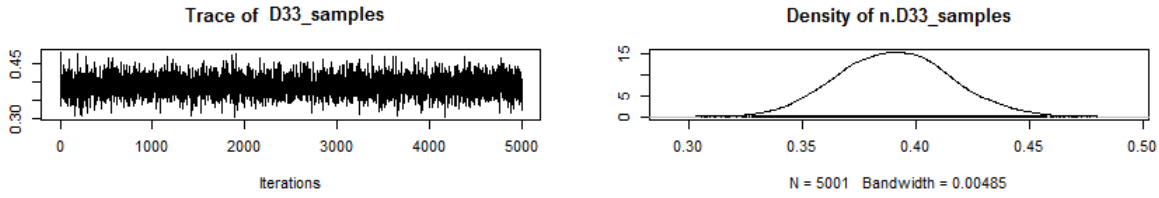

Figure B4.7: MCMC traces and posterior distribution plots for the parameter D_{33} in Model 2.