Pontificia Universidad Católica de Chile
Faculty of Mathematics
Department of Statistics

# New contributions to joint models of longitudinal and survival outcomes: Two-stage approaches

Valeria Leiva Yamaguchi

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR IN STATISTICS

August 2021

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

The undersigned hereby certify that they have read and recommend to the Faculty of Mathematics for acceptance a thesis entitled **New contributions to joint models of longitudinal and survival outcomes: Two-stage approaches** by **Valeria Leiva Yamaguchi** in partial fulfillment of the requirements for the degree of **Doctor in Statistics**.

Date: August, 2021

Research Supervisor: _____

Danilo Alvares
Pontificia Universidad Católica de Chile

Research Co-Supervisor: _____

Susana Eyheramendy
Universidad Adolfo Ibáñez

Examing Committee: _____

Cristian Meza
Universidad de Valparaíso

_____

Manuel Galea
Pontificia Universidad Católica de Chile

_____

Mauricio Castro
Pontificia Universidad Católica de Chile

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Date: August 2021

| | | |
|---|---|---|
| Author | : | Valeria Leiva Yamaguchi |
| Title | : | New contributions to joint models of longitudinal and survival outcomes: Two-stage approaches |
| Department | : | Statistics |
| Degree | : | Doctor in Statistics |
| Convocation | : | August |
| Year | : | 2021 |



Signature of Author

TO MY PARENTS

# Acknowledgements

Quiero agradecer a varias personas que estuvieron conmigo durante todo este proceso e hicieron que el camino fuese mucho más ameno. En primer lugar, quisiera agredecer a mi tutor Danilo Alvares por su apoyo, tiempo y confianza para realizar este proyecto de tesis. Gracias por tu paciencia y amistad.

Estoy en deuda con mi co-tutora Susana Eyheramendy por su ayuda y guía desde el primer día de este proceso. Gracias por todo el apoyo estadístico brindado durante todo este camino.

Quiero agradecer a todos los miembros administrativos del Departamento de Estadística por recibirme siempre con los brazos abiertos y una linda sonrisa. En particular, a mi Tinita, quién ha estado ahí desde el inicio de toda esta aventura. A todos los profesores, que de alguna manera han tenido una gran influencia en mi desarrollo como estadístico.

A mis compañeros de estudios por su disposición a discutir problemáticas estadísticas, pero sobre todo por hacer cada día más agradable. Mención honrosa a Angélica Domínguez, Darlin Soto, Eduardo Alarcón, Diego Morales y Alejandro Sepúlveda.

Quisiera agradecer al Núcleo Milenio Midas por el sorporte informático brindado durante la última etapa de este proceso. También me gustaría agradecer a la Vicerrectoría de Investigación (VRI) por el apoyo financiero otorgado durante todos estos años.

Finalmente, quisiera agradecer a mi familia y amigos por confiar en mis capacidades y siempre estar a mi lado en todo momento.

# Contents

# Abstract

Joint models of longitudinal and survival outcomes have gained much popularity over the last three decades. This type of modeling consists of two submodels, one longitudinal and one survival, which are connected by some common term. Unsurprisingly, sharing information makes the inferential process highly time-consuming. This problem can be overcome by estimating the parameters of each submodel separately, leading to a natural reduction in the complexity of joint models, but often producing biased estimates. Hence, we propose different two-stage strategies that first fits the longitudinal submodel and then plug the shared information into the survival submodel. Our proposals are developed for both the frequentist and Bayesian paradigms. Specifically, our frequentist two-stage approach is based on the simulation-extrapolation algorithm. On the other hand, we propose two Bayesian approaches, one inspired by frailty models and another that uses maximum a posteriori estimations and longitudinal likelihood to calculate posterior distributions of random effects and survival parameters. Based on simulation studies and real applications, we empirically compare our two-stage approaches with their main competitors. The results show that our methodologies are very promising, since they reduce the estimation bias compared to other two-stage methods and require less processing time than joint specification approaches.

# List of Tables

# List of Figures

# Chapter 1

# Introduction

In medical studies, it is common to find outcomes of interest of the following nature: the moment of occurrence of an event, such as clinical diagnosis, cure of a disease, or death; and an unobserved latent process measured repeatedly over time, as the indicator of progression for a certain disease. A classic example of such studies can be found in human immunodeficiency virus (HIV) clinical trials (Goldman et al., 1996), where patients affected by HIV were randomly assigned to receive one of two possible treatments. The evolution of the disease was monitored through the CD4 cell counts, which were measured over time. Typically, the time to death and the longitudinal trajectory of cell counts are modeled separately. Nevertheless, some researchers may potentially be interested in studying the association between both processes and therefore a joint modeling is required (Rizopoulos, 2012; Elashoff et al., 2016).

In a joint model specification, survival analysis incorporates into the risk of an event a biomarker measured repeatedly over time (endogenous covariate), which is usually mod-

eled as a longitudinal process (Ibrahim and Molenberghs, 2009). On the other hand, longitudinal studies model possible bias due to missing not at random data (Little and Rubin, 2019) through a survival process (Yu et al., 2004). In both cases, the joint modeling provides complete inference (longitudinal, survival, and association between both of them), reduces estimation bias, increases statistical efficiency, and conveniently makes predictions of outcomes (Muthén et al., 2009; Ibrahim et al., 2010; Wang et al., 2012). See Papageorgiou et al. (2019), Furgal et al. (2019), and Alsefri et al. (2020) for a review on joint models up to date.

These models have been extended in different directions with important practical applications, mainly in epidemiological research (Taylor et al., 2013). As one of the first proposals, Tsiatis et al. (1995) developed a two-stage procedure for this class of joint models. First, the longitudinal model is fitted. Then, the estimated values for the longitudinal trajectory are used as an endogenous covariate in a Cox proportional hazards model. The main advantage of this approach is that it avoids computational complexity problems by ignoring the potential joint nature of both processes. However, the main weakness of this methodology is that survival model parameter estimates are often biased (Faucett and Thomas, 1996; Tsiatis and Davidian, 2004; Ye et al., 2008; Rizopoulos, 2012).

Many authors have proposed two-stage alternatives that try to correct the estimation bias. For instance, Wulfsohn and Tsiatis (1997) estimated parameters within the Cox model while the longitudinal covariate is measured occasionally and with measurement error. They relaxed the distributional assumption of random effects to a distribution with a smooth density. Ye et al. (2008) and Albert and Shih (2010) proposed two-stage strategies based on regression calibration with longitudinal data modeled nonparametrially. Murawska et al. (2012) developed a two-stage approach using a non-linear mixed-effects

model and empirical Bayes estimates of subject-specific parameters as predictors in the Cox proportional-hazards model. Viviani et al. (2014) formulated an approach for non-Gaussian longitudinal data using an expansion of shared parameter models and assumed that the longitudinal trajectory up to the current time may have an impact on the hazard of dropout. Donnelly et al. (2018) considered a two-stage approach based on the Coxian phase-type distribution to represent the survival process. Huong et al. (2018) proposed a modified two-stage method considering an approximation of the fully joint log-likelihood function to estimate parameters from longitudinal process to later be incorporated into the survival model. In addition, some authors have developed two-stage model-based approaches for multivariate longitudinal and survival outcomes (Guler et al., 2017; Mauff et al., 2020).

Even so, these proposals are not generalized to different joint model structures. Thus, making the inference simultaneously is still the most recommended option, even though it is computationally more intense (Hsieh et al., 2006; Ye et al., 2008; Ibrahim et al., 2010; Andrinopoulou et al., 2018). However, the joint model specification has at least two drawbacks: *i)* identifiability problems due to the large number of parameters (Henderson et al., 2000; Wu, 2009; Gould et al., 2015; Wu and Yu, 2016; Papageorgiou et al., 2019); and *ii)* requirement for numerical integrations that can make the inferential process time-consuming (Lesaffre and Spiessens, 2001; Pinheiro and Chao, 2006; Rizopoulos et al., 2009; Wu et al., 2010; Barrett et al., 2015).

Two-stage approaches dodge both problems (Self and Pawitan, 1992; Tsiatis et al., 1995) but, as previously discussed, they often produce biased estimates and/or are not extended to any joint modeling. Hence, in order to overcome the potential limitations of traditional methods, we propose three two-stage strategies that aim to reduce or eliminate

the estimation bias with low processing times as well as providing easy-to-implement alternatives to more complex formulations of joint models. One of these proposals is developed within the context of frequentist statistics based on the simulation-extrapolation algorithm (Cook and Stefanski, 1994). In contrast, the other two proposals are based on the Bayesian paradigm, one inspired by frailty models (Ibrahim et al., 2001) and another that uses maximum a posteriori estimations and longitudinal likelihood to calculate posterior distributions of random effects and survival parameters.

This thesis is organized as follows: Chapter 2 introduces a general formulation of joint models for longitudinal and survival outcomes. Chapter 3 presents our frequentist two-stage proposal. Chapter 4 describes our two Bayesian proposals for two-stage estimation of joint models. Chapter 5 presents a brief description of three real datasets. Chapter 6 describes a simulation scheme of synthetic data from a joint modeling. Chapters 7 and 8 compare our methodologies against their main competitors using real and simulated data, respectively. Finally, Chapter 9 discusses the advantages, limitations, and extensions of the methods developed in this thesis.

# Chapter 2

# Joint formulation

Joint modeling for longitudinal and survival data has become very popular medical applications (Neuhaus et al., 2009; Wu et al., 2011). This class of models is a useful tool when it is necessary to study the association between repeated measurements and time until an event of interest (Papageorgiou et al., 2019). When considering both information together (simultaneously) into a single model, the estimates are less biased and there is an increase in statistical efficiency, since clinical hypotheses consider in advance that longitudinal and survival data are connected in some way (Muthén et al., 2009; Ibrahim et al., 2010). In the next sections, we will introduce the key elements for formulating a standard joint model.

## 2.1 Mixed-effects models for longitudinal data

Generally, mixed-effects models are considered for the estimation of the longitudinal process (Pinheiro and Bates, 2000). These models are based on the idea that each individual in

the population has his/her own subject-specific mean response profile over time. To introduce this representation of longitudinal data, we let $y_i(t)$ denote the response of individual $i$, $(i = 1, \ldots, n)$ at time $t$. A standard mixed-effect model has the following functional form:

$$\begin{aligned} y_i(t) &= \mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y) + \epsilon_i(t), \\ \epsilon_i(t) &\sim \text{Normal}(0, \sigma^2), \end{aligned} \tag{2.1}$$

where $\mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)$ denotes a trajectory function at time $t$, $\boldsymbol{\theta}_y$ is a parameter vector, and $\boldsymbol{b}_i$ represents an $M$-dimensional vector of individual random effects, often assumed normally distributed (Verbeke, 1997). Additionally, $\epsilon_i(t)$ denotes a random error term and follows a normal distribution with zero mean and variance $\sigma^2$. The trajectory function can be formulated in different ways, but the most common corresponds to the linear mixed model specification with $\mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\beta}) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i$, where $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i(t)$ are covariates of the individual $i$ at time $t$.

One of the main advantages of the mixed-effects model is to predict how individual response trajectories change over time. On the other hand, the main challenge for the analysis of longitudinal data is the missing data. Although in longitudinal studies, measurements are collected repeatedly over time in pre-specified follow-up times, it is often some individuals may fail to show up at a scheduled time for various reasons. Depending on the features of missing values patterns, we can distinguish two types of mechanisms: monotone and non-monotone missingness. The first type is when a patient leaves the study before its completion or when a individual does not provide information about initial response measurements, but appears later on and stays in the study until completion. On the other hand, the second type of missing data is when patients are missing at intermittent times, i.e., that other measurements are observed following missing values. The appropriateness of different procedures of analysis of incomplete longitudinal data is determined

by the missing data mechanism. In particular, there are three types of mechanisms (Little and Rubin, 2019):

- *Missing Completely at Random* (MCAR): occurs when the probability that the responses are missing is unrelated to the longitudinal outcome, meaning that is unrelated to both the specific values that individuals would have been obtained and the set of observed responses.
- *Missing at Random* (MAR): the probability of missingness depends on the set of observed responses, but it is unrelated to the longitudinal outcomes that would have been obtained.
- *Missing Not at Random* (MNAR): when the probability the longitudinal outcomes are missing depends on the set of responses observed and unobserved.

These missing data mechanisms may affect the covariates of longitudinal studies, which are defined as time-dependent and time-independent covariates. Time-dependent covariates are measured repeatedly over time and their values change at the same time. On the other hand, time-independent (or baseline) covariates are only measured once and their values do not change over time. In a mixed-effects model, when covariates have missing values, we must use a model for incompletely observed covariates in likelihood inference (Little, 1995).

## 2.2 Proportional hazard models for survival data

When we are interested in modeling the time to an event of interest occurs, Cox-type regression models are widely used (Cox, 1972). To introduce such type of model, we

denote $T_i^*$ as the true event time for individual $i$ and $C_i$ the censoring time, then $T_i = \min(T_i^*, C_i)$ is the observed event time and $\delta_i = \mathrm{I}(T_i^* \leq C_i)$ represents the event indicator. The proportional hazard models assume that covariates have a multiplicative effect on the hazard function for an event, and the formulation is given by:

$$h_i(t \mid \mathcal{M}_i(t)) = h_0(t) \exp\left\{\boldsymbol{w}_i^\top \boldsymbol{\gamma} + \alpha \mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)\right\}, \tag{2.2}$$

where $h_0(t)$ represents an arbitrary and left unspecified baseline hazard function at time $t$ and $\boldsymbol{w}_i$ is a covariate vector with coefficients $\boldsymbol{\gamma}$. $\mathcal{M}_i(t) = \{\mu_i(l \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y), 0 \leq l < t\}$ denotes the history of the longitudinal process up to $t$ and $\mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)$ is the underlying true trajectory from the longitudinal submodel (2.1), which has the role of connecting both processes while $\alpha$ measures the strength of this association. In order to simplify the notation, we will omit the term $\mathcal{M}_i(t)$ when specifying a hazard function. Under this approach, the distributional assumptions for $T_i^*$ are hidden in the specification of the baseline hazard. For instance, when assuming a Weibull specification for the survival submodel (2.2), the baseline hazard takes the form $h_0(t) = \phi\, t^{\phi-1} \exp(\gamma_0)$. When the interest is on the survival time and we wish to explain the effect of the longitudinal trajectory as a time-dependent covariate, traditional approaches to analyze the time until the event of interest are not applicable, because these assume that the time-dependent covariates are predictable processes and measured without error (Cox, 1975). In particular, the standard survival models require that the time-dependent covariates are external, i.e., that their value at a certain instant is not affected by the appearance of the event of interest. However, the type of time-dependent covariates found in longitudinal studies does not satisfy this condition, since they are the result of a stochastic process generated by the individual that is directly related to the failure mechanism (Kalbfleisch and Prentice, 2002).

The Cox model can be extended to time-dependent covariates, but however this type

of model works well only for exogenous time-varying covariates. For endogenous time-dependent covariates, such as biomarkers measured on patients, this approach can cause biases and poor coverage properties (Sweeting and Thompson, 2011). One of the main reasons is that it assumes that the level of the markers remains constant on the next visit.

## 2.3 Joint specification (JS) approach

Based on the mixed-effects (2.1) and proportional hazard (2.2) submodels, we introduce the standard joint model of longitudinal and survival data. The main idea behind this class of models is to couple the survival model for continuous time-to-event process with a mixed-effects model for the longitudinal outcome. The main estimation approach that has been proposed for joint modeling is the *maximum likelihood-based estimator* (Rizopoulos, 2012). The inferential methods that simultaneously consider the estimation of all parameters of a joint model depend on the specification of the joint density between both processes. Let $\boldsymbol{y}$ and $\boldsymbol{s}$ be the longitudinal and survival data, respectively. The vector of all parameters is specified by $\boldsymbol{\theta}$ and the random effects by $\boldsymbol{b}$. So, the joint density of $(\boldsymbol{y}, \boldsymbol{s})$ can be written as:

$$f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{\theta}) = \int f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta})\, f(\boldsymbol{b} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{b}, \tag{2.3}$$

where $f(\cdot)$'s denote density functions. There are different proposals for the specification of the conditional density $f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta})$ (Alvares, 2017). However, the most widely used approach is the *shared-parameter* specification (Wu and Carroll, 1988), in which it assumes that the longitudinal process is conditionally independent of the survival process given the shared information, i.e.:

$$f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}) = f(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\theta})\, f(\boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}), \tag{2.4}$$

where $f(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\theta})$ and $f(\boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta})$ are commonly specified according to submodels (2.1) and (2.2), respectively. Assuming that individuals are independent of each other, the likelihood function of $\boldsymbol{\theta}$ is expressed as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(\boldsymbol{y}_i, s_i \mid \boldsymbol{\theta}) \overset{(2.3)}{=} \prod_{i=1}^{n} \int f(\boldsymbol{y}_i, s_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{b}_i \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{b}_i \\ &\overset{(2.4)}{=} \prod_{i=1}^{n} \int f(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) f(s_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{b}_i \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{b}_i, \end{aligned} \tag{2.5}$$

where $f(\boldsymbol{b}_i \mid \boldsymbol{\theta})$ is typically assumed as a multivariate normal density with zero mean and $M \times M$ variance-covariance matrix $\boldsymbol{\Sigma}$. In some cases, the parameter vector $\boldsymbol{\theta}$ is more conveniently specified as $\boldsymbol{\theta}_y$ (longitudinal), $\boldsymbol{\theta}_s$ (survival), and $\boldsymbol{\theta}_b$ (random effects).

From the frequentist perspective, an expectation–maximization algorithm (Rizopoulos, 2012) robustly estimates the joint model parameters using the likelihood function (2.5). However, there are two potential limitations that deserve attention:

*i)* Joint models have many parameters to be estimated, so these parameters can become non-identifiable, i.e., two sets of different parameters can lead the same likelihood value (Henderson et al., 2000; Wu, 2009; Gould et al., 2015; Wu and Yu, 2016; Papageorgiou et al., 2019);

*ii)* Computational complexity can be challenging when the integral in (2.5) is high dimensional (Lesaffre and Spiessens, 2001; Pinheiro and Chao, 2006; Rizopoulos et al., 2009; Wu et al., 2010; Barrett et al., 2015).

## 2.4 Standard two-stage (STS) approach

Due to the potential complexity of numerical integration of the joint model likelihood function (2.5), several two-stage estimation methods for joint models of longitudinal and survival data have been proposed, both from the frequentist (Tsiatis et al., 1995; Wulfsohn and Tsiatis, 1997; Ye et al., 2008; Albert and Shih, 2010; Huong et al., 2018) and Bayesian (Murawska et al., 2012; Viviani et al., 2014; Mauff et al., 2020) perspectives. However, the most common two-stage procedure is described as follows:

- **Stage 1:** Fit the submodel (2.1), typically using a linear mixed model;
- **Stage 2:** Fit the submodel (2.2), typically using an extended Cox model, where the trajectory function estimated in **Stage 1** is considered as a time-varying covariate.

Note that this approach allows the specification of complex submodels while maintaining an acceptable processing time, since estimating the parameters of each submodel separately is computationally more stable. In addition, standard R-packages for each submodel can be used easily, for instance, **nlme** (Pinheiro and Bates, 2021) for the longitudinal submodel and **survival** (Therneau, 2011) for the survival submodel. However, as previously discussed, by ignoring the potential joint nature between both processes, this two-stage approach often produce biased estimates for the survival model regressors. (Faucett and Thomas, 1996; Tsiatis and Davidian, 2004; Ye et al., 2008; Rizopoulos, 2012).

# Chapter 3

# Frequentist perspective

This chapter is based on the article *"A novel two-stage bias correction approach for joint models of longitudinal and survival data"* (currently under review).

To get around the problems of both the joint specification (JS, see Section 2.3) and the standard two-stage (STS, see Section 2.4), we propose a novel two-stage (NTS) approach, based on a frequentist perspective, which prioritizes computational efficiency and robustness of parameter estimation. Like traditional two-stage approaches, our proposal starts by fitting the longitudinal submodel without considering survival data. In the second stage, we incorporate information from the longitudinal process to the survival process, using the simulation-extrapolation (SIMEX) methodology to estimate the survival model parameters. The step-by-step construction of our proposal and an algorithm that summarizes it are presented in the next sections.

## 3.1 Novel two-stage (NTS) approach

The general idea behind the SIMEX is to fix the bias of a particular estimator when the bias is due solely by the presence of a covariate measured with known errors. The standard SIMEX algorithm consists on three steps: *simulation*, *estimation*, and *extrapolation* (Cook and Stefanski, 1994). In the simulation step, the covariate measured with error is generated many times with some fixed and larger amount of measurement errors. From each of the generated covariate a new parameter estimator is obtained, and then a function is learned that relates the amount of measurement error with the parameter estimator. This function is then extrapolated to zero variability to obtained at last the presumably unbiased estimator.

Our proposal is strongly inspired by SIMEX steps. However, the joint models framework requires some relevant adaptations, such as the initialization step introduced next.

### 3.1.1 Initialization

Without loss of generality, we assume an additive measurement error specification (Carroll et al., 2006) for the longitudinal outcome of the individual $i$ at time $t$ given by:

$$y_i(t) = y_i^*(t) + \xi_i(t), \tag{3.1}$$

where $y_i(t)$ represents the observed error-prone value, $y_i^*(t)$ is the true unobserved value, and $\xi_i(t) \sim \text{Normal}(0, \hat{\sigma}_\xi^2)$. The term $\hat{\sigma}_\xi^2$ is called *measurement error variance* and it is assumed known (Yang et al., 2019). In practice, we recommend that the user set the value of $\hat{\sigma}_\xi^2$. However, we are aware that the scale of this variance is strictly associated with the longitudinal outcome magnitude and therefore setting it can be a difficult task. For instance, a small value can lead to results quite similar to the standard two-stage approach,

while a large value can inflate this variance causing an increase in the estimation bias. As an alternative, based on simulation studies, we propose to estimate the measurement error standard deviation as $\hat{\sigma}_\xi = p\,\hat{\sigma}$, where $p \in [0, 1]$ is chosen by the user and $\hat{\sigma}$ represents the error standard deviation of the longitudinal submodel (2.1) estimated from the observed longitudinal outcomes. In short, $\hat{\sigma}_\xi$ is defined as a fraction of $\hat{\sigma}$.

### 3.1.2 Simulation

Since the measurement error variance $\hat{\sigma}^2_\xi$ has been set, two tuning parameters of the SIMEX simulation step must be defined:

- $K$: number of simulations/repetitions. A large value of $K$ lead to a better SIMEX estimator according to the Monte Carlo error reduction criterion (Hastings, 1970). Usually, $K = 50$ or $K = 100$ works well (Cook and Stefanski, 1994).
- $\Delta = [\lambda_1, \ldots, \lambda_m]$ with $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_m$: grid of values used to create response variables with increasingly larger amounts of measurement error. Typically, $\Delta = [0.5, 1, 1.5, 2]$ is considered the default specification (Hardin et al., 2003; Carroll et al., 2006).

For each $k = 1, \ldots, K$ and $\lambda \in \Delta$, we generate "new" longitudinal outcomes with the following perturbation/contamination:

$$y_i(t; k, \lambda) = y_i(t) + \lambda^{1/2}\xi_i(t; k) \;\; \forall i, \tag{3.2}$$

where $\xi_i(t; k) \sim \text{Normal}(0, \hat{\sigma}^2_\xi)$. Note that we are simulating values from a normal distribution centered on $y_i(t)$ and with variance $\lambda\,\hat{\sigma}^2_\xi$. In particular, the total measurement error

variance for the $k$th simulated response variable, given a $\lambda$, is

$$\text{Var}\big[\boldsymbol{Y}(t;k,\lambda)\big] = \text{Var}\big[\boldsymbol{Y}(t)\big] + \lambda\text{Var}\big[\boldsymbol{\xi}(t;k)\big] = \hat{\sigma}^2_\xi\,\mathbf{I} + \lambda\,\hat{\sigma}^2_\xi\,\mathbf{I} = (1+\lambda)\,\hat{\sigma}^2_\xi\,\mathbf{I}, \quad (3.3)$$

where $\boldsymbol{Y}(t;k,\lambda)$, $\boldsymbol{Y}(t)$ and $\boldsymbol{\xi}(t;k)$ represent random vectors of dimension $n$ (number of individuals) and $\mathbf{I}$ is the $n \times n$ identity matrix.

Using $y_i(t;k,\lambda)$ instead of $y_i(t)$ for $i = 1,\ldots,n$, we estimate the trajectory function at time $t$ of the longitudinal submodel (2.1) and then connect it into the survival submodel (2.2), the same way as in STS approach (see Section 2.4). For each iteration, the estimated survival model parameters (let's say $\hat{\boldsymbol{\theta}}_s(k,\lambda)$) are stored as well as other potential quantities of interest.

### 3.1.3 Estimation

The average value of the estimated survival model parameters for each $\lambda$ is calculated:

$$\hat{\boldsymbol{\theta}}_s(\lambda) = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_s(k,\lambda). \quad (3.4)$$

The average value of other quantities of interest can also be obtained, e.g., estimated variance and standard error of each parameter (Apanasovich et al., 2009; He et al., 2012).

### 3.1.4 Extrapolation

The average estimates $\hat{\boldsymbol{\theta}}_s(\lambda_1),\ldots,\hat{\boldsymbol{\theta}}_s(\lambda_m)$ generate a trajectory of values for each survival model parameter according to the (sequential) increment in the variance $(1+\lambda)\,\hat{\sigma}^2_\xi\,\mathbf{I}$. Note that $\lambda = 0$ is equivalent to estimating the parameters using the original data (STS

approach), i.e., without applying our bias correction proposal. In SIMEX literature, this estimate is known as *naive estimation* (Carroll et al., 2006).

The trajectory of each estimated average parameter can be modeled using a polynomial or nonlinear regression (Cook and Stefanski, 1994; Stefanski and Cook, 1995). In general, a quadratic regression satisfactorily fits these values (Carroll and Küchenhoff, 1995; He et al., 2012) and therefore it is implemented in this thesis. Figure 3.1 illustrates an example for the extrapolation step using a quadratic regression.

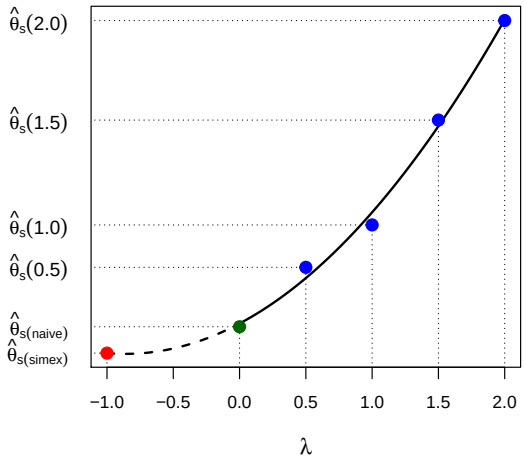


Figure 3.1: Illustration of a quadratic extrapolation function to estimate a (univariate) parameter $\theta_s$. The green and red points indicate the naive (STS approach) and SIMEX (NTS approach) estimators, respectively. The blue points are the estimates when increasing the measurement error variance, $(1 + \lambda)\,\hat{\sigma}^2_\xi$, for $\lambda = 0.5, 1, 1.5, 2$.

It is worth noting that the case of no measurement error is obtained when the variance in Equation (3.3) is equal to zero. Thus, the SIMEX estimator is calculated by extrapolat-

ing a polynomial or nonlinear regression back to $\lambda = -1$. Our proposal is summarized in Algorithm 1.

---

**Algorithm 1:** Novel two-stage (NTS) approach for joint models.

---

0 INITIALIZATION: Let $\hat{\sigma}_\xi^2$ be the measurement error variance defined either by the user directly or by specifying $p \in [0, 1]$ (see Section 3.1.1).

1 SIMULATION:

**for** $(k = 1 : K)$ **do**

**for** $(\lambda \in \Delta)$ **do**

– Generate

$$y_i(t; k, \lambda) = y_i(t) + \lambda^{1/2}\xi_i(t; k) \;\; \forall i,$$

where $\xi_i(t; k) \sim \text{Normal}(0, \hat{\sigma}_\xi^2)$.

– Estimate $\mu_i(t; k, \lambda)$ from (2.1) using $y_i(t; k, \lambda)$ instead of $y_i(t)$ $\forall i$.

– Estimate $\boldsymbol{\theta}_s(k, \lambda)$ from (2.2) using $\mu_i(t; k, \lambda)$ instead of $\mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)$ $\forall i$.

**end**

**end**

2 ESTIMATION: Obtain $\hat{\boldsymbol{\theta}}_s(\lambda)$ as the mean of the $K$ estimates of $\boldsymbol{\theta}_s(k, \lambda)$.

3 EXTRAPOLATION: Fit curve to the generated pairs $\big(\lambda, \hat{\boldsymbol{\theta}}_s(\lambda)\big)$. The extrapolated value at $\lambda = -1$ is the SIMEX estimator of $\boldsymbol{\theta}_s$.

---

# Chapter 4

# Bayesian approach

This chapter is based on articles *"A two-stage approach for Bayesian joint models of longitudinal and survival data: correcting bias with informative prior"* (Entropy, 23(1), 1-10, 2021) and *"A two-stage approach for Bayesian joint models: reducing complexity while maintaining accuracy"* (currently under review).

As we have discussed throughout this thesis, share information between submodels makes the inferential process highly time-consuming. From a Bayesian perspective, Markov chain Monte Carlo (MCMC) methods are computationally very demanding, especially for complex models with random effects (Gelman et al., 2013), which is the case for most joint models of longitudinal and survival data. Again, this problem can be alleviated by estimating the parameters of each submodel separately, leading to a natural reduction in the complexity of the joint modeling, but often producing biased estimates.

Since the philosophy of Bayesian statistics is different from the frequentist perspective, the following sections introduce the Bayesian version for joint specification (JS) and

standard two-stage (STS) approaches. Next, we will present our two Bayesian two-stage proposals for joint models.

## 4.1 Bayesian joint specification (BJS) approach

Let $\boldsymbol{y}$, $\boldsymbol{s}$ and $\boldsymbol{b}$ be the longitudinal data, survival data and random effects, respectively, and the vector of all parameters, $\boldsymbol{\theta}$, can be partitioned by random component, such as $\boldsymbol{\theta}_y$, $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_b$. Bayesian joint models for longitudinal and survival outcomes usually assume a full joint probability density given by (Armero et al., 2018; Alvares et al., 2021):

$$f(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{b}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s, \boldsymbol{\theta}_b) = f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s) f(\boldsymbol{b} \mid \boldsymbol{\theta}_b) \pi(\boldsymbol{\theta}_y) \pi(\boldsymbol{\theta}_s) \pi(\boldsymbol{\theta}_b). \tag{4.1}$$

The factors on the right hand side of (4.1) are the conditional joint density of the processes $\boldsymbol{y}$ and $\boldsymbol{s}$ given $\boldsymbol{b}$, $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_s$, $f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s)$; the conditional density of $\boldsymbol{b}$ given $\boldsymbol{\theta}_b$, $f(\boldsymbol{b} \mid \boldsymbol{\theta}_b)$; and the (independent) prior distributions of $\boldsymbol{\theta}_y$, $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_b$, $\pi(\boldsymbol{\theta}_y)$, $\pi(\boldsymbol{\theta}_s)$ and $\pi(\boldsymbol{\theta}_b)$, respectively. Similar to Equation (2.4), $f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s)$ can be factorized as $f(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\theta}_y) f(\boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s)$ (see Section 2.3 for more details).

From a joint specification approach, the posterior distribution $\pi(\boldsymbol{b}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s, \boldsymbol{\theta}_b \mid \boldsymbol{y}, \boldsymbol{s})$ is proportional to (4.1). This posterior distribution, partially derived from submodels (2.1) and (2.2), is highly complex due to the large number of parameters and random effects as well as potential integrations with no closed-form. Hence, as expected, the processing time is quite demanding and, in some cases, the Bayesian approach may be impractical (Ye et al., 2008; Andrinopoulou et al., 2018).

## 4.2 Bayesian standard two-stage (BSTS) approach

As we have discussed previously, two-stage strategies are very useful for reducing the complexity of joint models and speeding up the inferential process. From a frequentist point of view, Tsiatis et al. (1995) proposed one of the most popular two-stage approaches (see Section 2.4 for more details). The first stage is to fit the longitudinal submodel (2.1) and then the trajectory function $\mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)$ is calculated using the estimated parameters and random effects. Hence, this estimated trajectory function is considered as an endogenous time-varying covariate when fitting the survival submodel (2.2).

As a potential competitor, we use the Tsiatis et al. (1995) approach adapted to the Bayesian framework. Specifically, in the first stage, we calculate the *maximum a posteriori* (MAP) of parameters $\boldsymbol{\theta}_y$ and random effects $\boldsymbol{b}$ from longitudinal submodel fitted separately. Posterior mean or median can also be used instead. In the second stage, we incorporate the trajectory function into the survival submodel considering $\hat{\mu}_i(t \mid \hat{\boldsymbol{b}}_i, \hat{\boldsymbol{\theta}}_y)$, for $i = 1, \ldots, n$, where $\hat{\boldsymbol{b}}_i$ and $\hat{\boldsymbol{\theta}}_y$ are the MAP of $\boldsymbol{b}_i$ and $\boldsymbol{\theta}_y$, respectively, and then the posterior distribution of $\boldsymbol{\theta}_s$ is calculated via MCMC.

## 4.3 First novel Bayesian two-stage (NBTS1) approach

The first part of our two-stage proposal is similar to the Bayesian standard approach, i.e., the trajectory function $\mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)$ from the longitudinal submodel (2.1) is estimated using the MAP of $\boldsymbol{b}_i$ and $\boldsymbol{\theta}_y$. However, we propose the following modification to the survival submodel:

$$h_i(t) = \vartheta_i \, h_0(t) \exp\left\{ \boldsymbol{w}_i^\top \boldsymbol{\gamma} + \alpha \hat{\mu}_i(t \mid \hat{\boldsymbol{b}}_i, \hat{\boldsymbol{\theta}}_y) \right\}, \tag{4.2}$$

where $\vartheta_i$ denotes a multiplicative fixed-effect for individual $i$.

The role of $\vartheta_i$ is essential to satisfactorily correct the estimation bias by ignoring the potential joint nature between both processes. Specifically, what we propose is a very small perturbation using the individual fixed-effect. To do so, we specify a highly informative prior distribution for $\vartheta_i$, given by:

$$\vartheta_i \sim \text{Gamma}(\eta, \eta), \tag{4.3}$$

where $\eta$ is a known parameter and must be specified such that $\text{E}(\vartheta_i) = 1$ and $\text{Var}(\vartheta_i)$ small. Interpretatively, if $\vartheta_i$ is not perturbed (i.e., $\text{Var}(\vartheta_i) = 0$), then we turn to the Bayesian standard two-stage approach presented in Section 4.2. Moreover, note that if we assume that $\eta$ is an unknown parameter and so a hyperprior should be set for it, then the specification (4.2) becomes a Bayesian frailty model (Ibrahim et al., 2001).

## 4.4 Second novel Bayesian two-stage (NBTS2) approach

Like all two-stage methods for joint models, the first stage is to fit the longitudinal submodel separately. So, we calculate the MAP of the longitudinal and random effects parameters, which will be denoted by $\hat{\boldsymbol{\theta}}_y$ and $\hat{\boldsymbol{\theta}}_b$, respectively. Note that it is not necessary to estimate the random effects $\boldsymbol{b}$ and therefore, we can run a Bayesian estimation algorithm using the marginalized likelihood function (integrating out the random effects) of the longitudinal submodel (2.1).

In the second stage, we estimate the random effects $\boldsymbol{b}$ and survival parameters $\boldsymbol{\theta}_s$

replacing the full joint probability distribution (4.1) with:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{b}, \boldsymbol{\theta}_s \mid \hat{\boldsymbol{\theta}}_y, \hat{\boldsymbol{\theta}}_b) &= f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \hat{\boldsymbol{\theta}}_y, \boldsymbol{\theta}_s) f(\boldsymbol{b} \mid \hat{\boldsymbol{\theta}}_b) \pi(\boldsymbol{\theta}_s) \\ &\overset{(2.4)}{=} f(\boldsymbol{y} \mid \boldsymbol{b}, \hat{\boldsymbol{\theta}}_y) f(\boldsymbol{s} \mid \boldsymbol{b}, \hat{\boldsymbol{\theta}}_y, \boldsymbol{\theta}_s) f(\boldsymbol{b} \mid \hat{\boldsymbol{\theta}}_b) \pi(\boldsymbol{\theta}_s), \end{aligned} \tag{4.4}$$

where $f(\cdot)$'s and $\pi(\cdot)$ have the same functional form as in (4.1) and (2.4), except that the parameters $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_b$ are now assumed to be known and so it is not necessary to include their prior distributions in (4.4). Note that when using $\boldsymbol{\theta}_b$ estimated from the longitudinal submodel (i.e., $\hat{\boldsymbol{\theta}}_b$), random effects $\boldsymbol{b}$ become fixed effects, further reducing the complexity of (4.4) with respect to the full joint probability distribution (4.1).

Essentially, our proposal simplifies/approximates the likelihood function from the joint approach by assuming that the shared parameters can be satisfactorily estimated from their MAP using the longitudinal submodel separately. Proposition 4.4.1 presents asymptotic aspects that support our proposal.

**Proposition 4.4.1.** *For fixed sample size $n$, and as the number $n_i$ of repeated measurements per individual in the longitudinal process $\boldsymbol{y}$ increases, the maximum a posteriori (MAP) of $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_s$ from the Bayesian joint specification (BJS) and the second novel Bayesian two-stage (NBTS2) approaches will closely resemble.*

*Proof.* Let $f(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)$ and $f(s_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s)$ be the longitudinal and survival likelihood functions for individual $i$, respectively, where $\boldsymbol{y}_i = \{y_{it_{ij}} : t_{ij} = 1, \ldots, n_i\}$. We will analyze the behavior of the *maximum a posteriori* (MAP) of $\boldsymbol{\theta}_y$ from the conditional joint

log-likelihood function of individual $i$, as $n_i$ grows:

$$\begin{aligned} l_{n_i}(\boldsymbol{\theta}_y) &= \log\left(\prod_{j=1}^{n_i} f(y_{it_{ij}} \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)\right) + \log\big(f(s_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s)\big) \\ &= \sum_{j=1}^{n_i} \log\big(f(y_{it_{ij}} \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)\big) + \log\big(f(s_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s)\big) \\ &= -\frac{n_i}{2}\log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2}\sum_{j=1}^{n_i}\big(y_{it_{ij}} - \mu_i(t_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)\big)^2 + \log\big(f(s_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y, \boldsymbol{\theta}_s)\big). \end{aligned} \tag{4.5}$$

Calculating $\lim_{n_i\to\infty}\big[l_{n_i}(\boldsymbol{\theta}_y)/n_i\big]$, we note that the longitudinal term is Op(1), while the survival one is op(1). Extending this result to the joint likelihood for all individuals is straightforward. In summary, as $n_i$ grows, the MAP of $\boldsymbol{\theta}_y$ depends only on the longitudinal submodel. In this sense, we can calculate the MAP of $\boldsymbol{\theta}_y$ both from the BJS and from the NBTS2 first stage equivalently.

Taking the MAP of $\boldsymbol{\theta}_y$ (let's say $\hat{\boldsymbol{\theta}}_y$) from the NBTS2 approach, our next goal is to calculate the MAP of $\boldsymbol{\theta}_s$. Hence, similar to a profiled likelihood strategy (Murphy and van der Vaart, 2000), we can consider $\hat{\boldsymbol{\theta}}_y$ as a vector of nuisance parameters and then estimate the MAP of $\boldsymbol{\theta}_s$ from the the conditional joint log-likelihood function:

$$l(\boldsymbol{\theta}_s) = \log\big(f(\boldsymbol{y} \mid \boldsymbol{b}, \hat{\boldsymbol{\theta}}_y)\big) + \log\big(f(\boldsymbol{s} \mid \boldsymbol{b}, \hat{\boldsymbol{\theta}}_y, \boldsymbol{\theta}_s)\big). \tag{4.6}$$

The MAP of $(\boldsymbol{\theta}_y, \boldsymbol{\theta}_s)$ jointly estimated from the conditional joint log-likelihood function is equivalent to calculating the MAP of $\boldsymbol{\theta}_y$ from $l_{n_i}(\boldsymbol{\theta}_y)$ (assuming that $n_i \to \infty$) and then obtaining the MAP of $\boldsymbol{\theta}_s$ from $l(\boldsymbol{\theta}_s)$. Note that Equation (4.6) is specified as the log-likelihood function in NBTS2 second stage.

So far we have omitted the specification of random effects and priors for the estimation of the MAP of $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_s$. We assume that the distribution of $\boldsymbol{b}$, $f(\boldsymbol{b} \mid \boldsymbol{\theta}_b)$, from the BJS

approach is the same as the prior distribution of fixed effects $\boldsymbol{b}$, $\pi(\boldsymbol{b} \mid \hat{\boldsymbol{\theta}}_b)$, from the NBTS2 approach. So, the contribution of random effects to calculate the MAP of $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_s$ from both approaches is the same. Additionally, even if the random effects distributions are misspecified, Rizopoulos et al. (2008)'s Theorem 1 demonstrates that, as $n_i$ grows, the estimates of $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_s$ will be minimally affected. On the other hand, we assume that the prior distributions are proper and weakly-informative, so they must have an irrelevant role in calculating the MAP of $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_s$. Furthermore, as the sample size increases, the choice of prior distributions have a minimal impact. □

## 4.5 Prior distributions

To complete the Bayesian models formulation, we need to assign prior distributions to all parameters and hyperparameters. As a standard specification, we assume independent and diffuse marginal prior distributions, i.e., proper distributions with a large variance (Gelman et al., 2013). More specifically, all longitudinal and survival regression coefficients (including the association parameter) follow normal distributions with zero mean and large variance; $\sigma$ follows a weakly-informative half-Cauchy$(0, 5)$ (Gelman, 2006); and $\boldsymbol{\Sigma}$ follows an inverse-Wishart$(V, r)$, where $V$ is an $M \times M$ identity matrix, $r = M$ is the degrees-of-freedom parameter (Schuurman et al., 2016). Once the baseline hazard function $h_0(t \mid \boldsymbol{\theta}_s)$ is defined, diffuse priors are also specified for its parameters.

# Chapter 5

# Datasets

We introduce three real datasets, which are available in the R-package **JM** (Rizopoulos, 2018), in order to further compare the performance of the methods presented in this thesis.

## 5.1 Liver cirrhosis (prothro) dataset

This dataset includes 488 patients with histologically verified cirrhosis, with 237 of them randomized to treatment with prednisone and the remaining received placebo (Andersen et al., 1993). Liver cirrohosis is a generic term that includes all forms of liver disease in which healthy liver tissue is replaced with scar tissue and the liver is permanently damaged. This study took place from 1962 to 1974 in Copenhagen, and the principal objective was to evaluate whether prednisone prolongs survival for patients with cirrhosis. The visits are specified at 3, 6, and 12 months, and yearly thereafter, and provide records for several clinical and biochemical variables. Figure 5.1 shows subject-specific longitudinal trajec-

tories of the prothrombin index (in logarithmic scale) and the Kaplan-Meier estimate for the time to death by treatment.

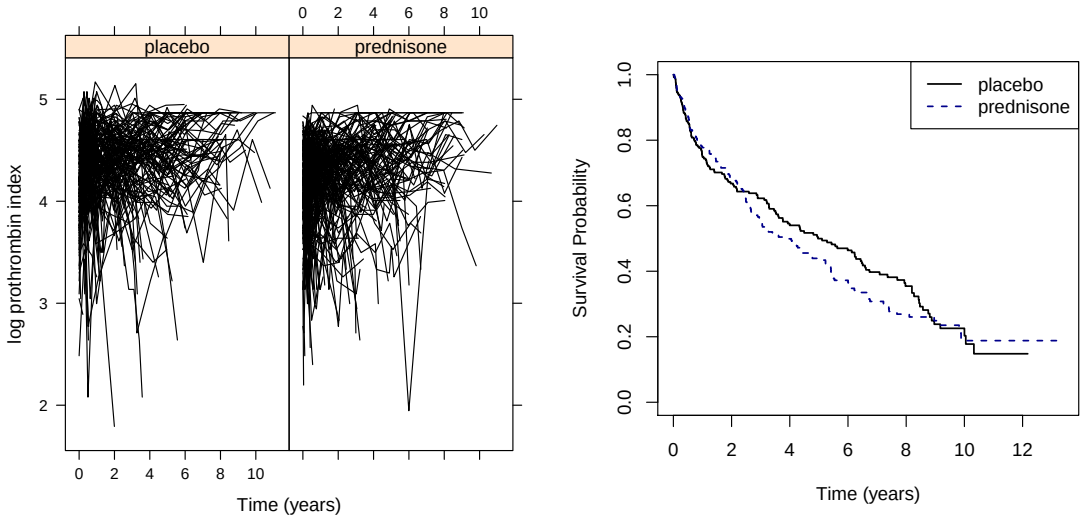


Figure 5.1: Left panel: Longitudinal trajectories of the prothrombin index (in logarithmic scale) by treatment. Right panel: Kaplan-Meier estimate for the time to death by treatment.

## 5.2 Acquired immunodeficiency syndrome (aids) dataset

This study consists of 467 patients with advanced human immunodeficiency virus infection during antiretroviral treatment who had failed or were intolerant to zidovudine therapy (Goldman et al., 1996). The principal aim was to compare the efficacy of two antiretroviral drugs/treatments, namely didanosine (ddI) and zalcitabine (ddC), in the time to death. Patients were randomly assigned to receive either ddI or ddC, and the evolution of the disease was monitored through the CD4 cell counts were recorded at study entry, for these

patients specified visits at 2, 6, 12, and 18 months thereafter. Figure 5.2 shows longitudinal trajectories of the square root of the CD4 cell counts and the Kaplan-Meier estimate for the time to death by treatment.

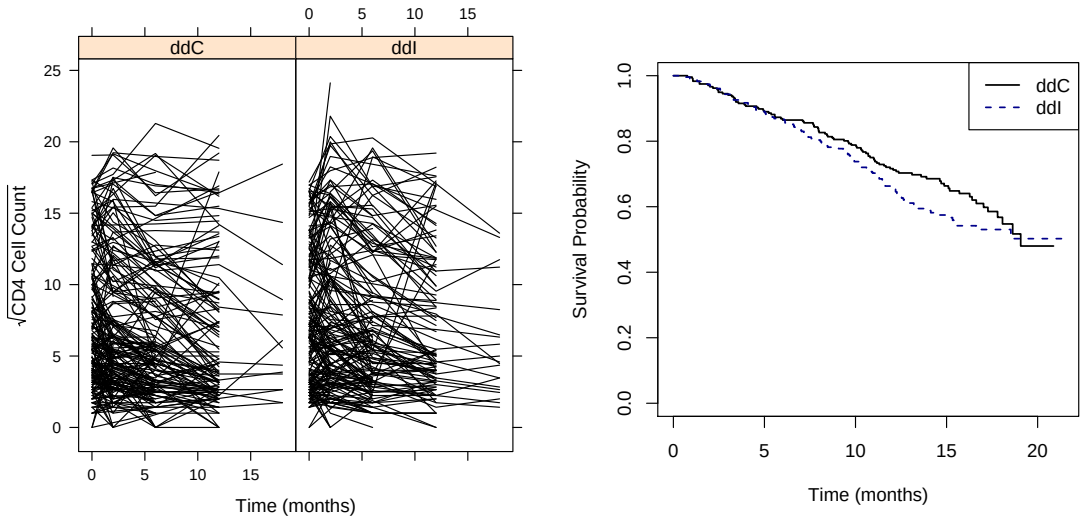


Figure 5.2: Left panel: Longitudinal trajectories of the square root CD4 cell counts by treatment. Right panel: Kaplan-Meier estimate for the time to death by treatment.

## 5.3 Primary biliary cirrhosis (pbc) dataset

This dataset includes 312 patients with primary biliary cirrhosis, a rare autoimmune liver disease, at Mayo Clinic trial conducted between 1974 and 1984, where 158 patients were randomized to D-penicillamine and 154 to placebo (Murtaugh et al., 1994). The outcome of primary interest was patient survival and whether this could be prolonged by D-penicillamine. Figure 5.3 shows longitudinal profiles of the serum bilirubin (in logarithmic scale) and Kaplan-Meier estimate for the time to death by treatment.

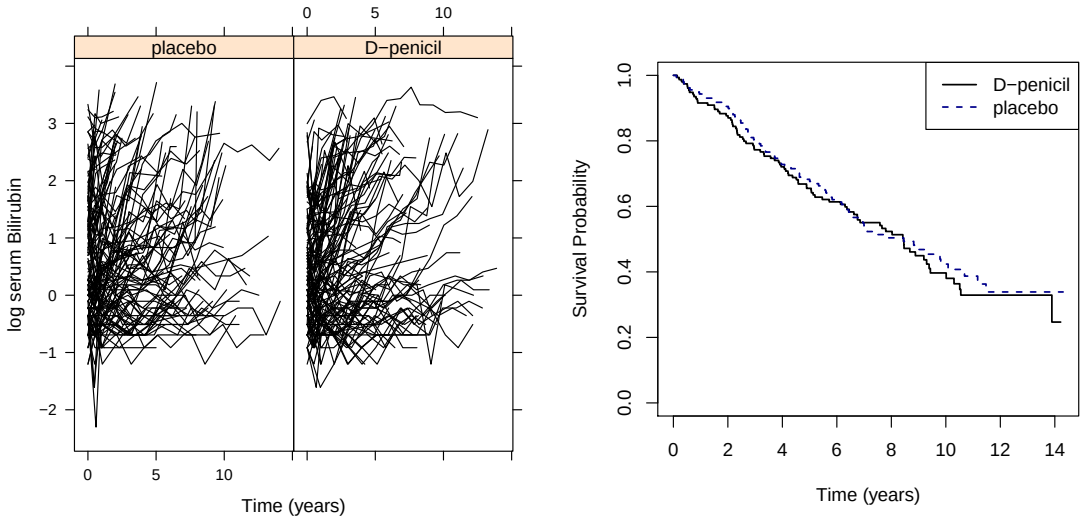


Figure 5.3: Left panel: Longitudinal trajectories of the serum bilirubin (in logarithmic scale) by treatment. Right panel: Kaplan-Meier estimate for the time to death by treatment.

# Chapter 6

# Simulating data from joint models

Chapter 8 will compare the performance of the methodologies presented in this thesis regarding the dual problem "biased estimation versus high computational time". To do so, a joint model simulation scheme should be used. Therefore, we develop in this chapter a simulation algorithm for joint models based on (2.1) and (2.2). More specifically, the longitudinal process for an individual $i$ at time $t$ is written as a linear mixed model:

$$\begin{aligned} y_i(t) &\sim \text{Normal}\big(\mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y), \sigma^2\big), \\ \mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y) &= \beta_0 + b_{0i} + (\beta_1 + b_{1i})t + \beta_2 x_i, \\ \boldsymbol{b}_i &= (b_{0i}, b_{1i})^\top \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{6.1}$$

where $\boldsymbol{\theta}_y = (\beta_0, \beta_1, \beta_2, \sigma^2)^\top$ and $\boldsymbol{\theta}_b = \boldsymbol{\Sigma}$. The covariate $x_i$ is a binary group indicator simulated from a Bernoulli distribution with probability $0.5$. The random effects and measurement error are independent.

The hazard function of the survival time $T_i$ of individual $i$ is specified as follows:

$$h_i(t) = \phi\, t^{\phi-1} \exp\left\{\gamma_0 + \gamma_1 x_i + \alpha \mu_i(t \mid \boldsymbol{b}_i, \boldsymbol{\theta}_y)\right\}, \tag{6.2}$$

where $h_0(t)$ is a Weibull baseline hazard with $\phi$ and $\exp(\gamma_0)$ being shape and scale parameters, respectively, and $\boldsymbol{\theta}_s = (\phi, \gamma_0, \gamma_1, \alpha)^\top$, where $\gamma_1$ and $\alpha$ are group and association parameters, respectively.

As a preliminary simulation step, all parameters ($\boldsymbol{\theta}_y$, $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_b$), number of individuals ($n$), minimum number of longitudinal observations ($m_{\min}$), and maximum observational time ($t_{\max}$) must be set. Then, the covariate $x_i$ and the random effects $\boldsymbol{b}_i$, for $i = 1, \ldots, n$, are simulated.

The event time for individual $i$, $T_i^*$, is obtained by simulating a value $u_i$ from the standard uniform distribution and solving the following equation:

$$u_i = S_i(t) = \exp\left\{-\int_0^{T_i^*} h_i(w)\,\mathrm{d}w\right\} \iff \log(u_i) + \int_0^{T_i^*} h_i(w)\,\mathrm{d}w = 0. \qquad (6.3)$$

Note that calculating $T_i^*$ is a (univariate) root-finding problem, which can be solved using numerical integration techniques such as Gauss-type quadrature rules (Crowther and Lambert, 2013).

The censoring time for each individual, $C_i$, is generated from a uniform distribution on the interval $[0, t_{\max}]$ and then the observed time is set as $T_i = \min\{T_i^*, C_i\}$ and the event indicator as $\delta_i = \mathrm{I}(T_i^* \leq C_i)$.

The number of longitudinal observations of individual $i$, $n_i$, is set as $m_{\min}$ plus the largest integer less than $T_i$ (i.e., $\lfloor T_i \rfloor$). The recording times of the repeated measurements are equispaced set from 0 to $\lfloor T_i \rfloor$. The random errors $\epsilon_i(t_1), \ldots, \epsilon_i(t_{n_i})$ are simulated from a normal distribution with zero mean and variance $\sigma^2$. Finally, the longitudinal observations of individual $i$, $y_i(t_1), \ldots, y_i(t_{n_i})$, are computed according to Equation (6.1).

The simulation scheme to jointly generate longitudinal and survival data is summarized in Algorithm 2.

**Algorithm 2:** Simulation scheme for joint models.

0 INITIALIZATION: Set $\boldsymbol{\theta}_y$, $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$, $n$, $m_{\min}$, and $t_{\max}$.

1 SIMULATING SURVIVAL DATA:

– Simulate $x_i \sim$ Bernoulli$(0.5)$ and $\boldsymbol{b}_i \sim$ Normal$(\mathbf{0}, \boldsymbol{\Sigma})$ $\forall i$.

– Calculate $T_i^*$ $\forall i$ based on the survival submodel (6.2) using (6.3).

– Simulate $C_i \sim$ Uniform$(0, t_{\max})$ $\forall i$.

– Set $T_i = \min\{T_i^*, C_i\}$ and $\delta_i = \mathrm{I}(T_i^* \leq C_i)$ $\forall i$.

2 SIMULATING LONGITUDINAL DATA:

– Set $n_i = m_{\min} + \lfloor T_i \rfloor$ $\forall i$.

– Set $0 = t_1, \ldots, t_{n_i} = \lfloor T_i \rfloor$ $\forall i$ equispaced.

– Simulate $\epsilon_i(t) \sim$ Normal$(0, \sigma^2)$, $t = t_1, \ldots, t_{n_i}$ $\forall i$.

– Compute $y_i(t_1), \ldots, y_i(t_{n_i})$ $\forall i$ based on the longitudinal submodel (6.1).

# Chapter 7

# Application

We compared the inferential procedures, both frequentist and Bayesian, presented in this thesis using datasets introduced in Chapter 5. The three datasets contain two observation formats, the long one for longitudinal data and the short one for survival data. For all analyses, we used the joint formulation based on (6.1) and (6.2), where the longitudinal submodel is specified as a linear mixed model with random intercept and slope, and the survival submodel is defined as a proportional hazard model with Weibull specification. Moreover, $x_i$ is the treatment/drug indicator variable for each dataset. Finally, for Bayesian models, prior distributions were specified as in Section 4.5.

For the novel two-stage (NTS) approach from a frequentist perspective, we fixed the number of simulations/repetitions and the grid of values used to create response variables as recommended in the literature, i.e., $K = 50$ and $\Delta = [\lambda_1, \ldots, \lambda_4] = [0.5, 1, 1.5, 2]$, respectively (see Section 3.1.2 for more details). The parameter $p \in [0, 1]$, which defines the measurement error standard deviation, was set according to a simulation study sensitivity

analysis presented in Appendix A.1. Specifically, $p = 0.9$ was used for the prothro and aids datasets, and $p = 0.1$ was used for the pbc dataset.

For the first novel Bayesian two-stage (NBTS1) approach, we set the individual multiplicative fixed-effect parameter, $\eta$, based on a simulation study sensitivity analysis presented in Appendix B.1 (see Section 4.3 for more details). Specifically, $\eta = 10, 20, 30$ were used for the prothro, aids and pbc datasets, respectively.

The MCMC configuration for all Bayesian models was defined as follows: 2000 iterations with warm-up of 1000 for the joint model using the joint specification (BJS) and for the longitudinal submodel from the novel two-stage (NBTS1 and NBTS2) and standard two-stage (BSTS) approaches. Additionally, 1000 iterations with warm-up of 500 were set to run the survival submodel from the three Bayesian two-stage approaches. All models were implemented using the R-package **rstan** (Stan Development Team, 2020).

Table 7.1 shows the comparative results for the three frequentist approaches (JS, NTS, and STS) and the four Bayesian approaches (BJS, NBTS1, NBTS2, and BSTS) for each dataset. In particular, estimation comparisons are focused on the group and association parameters, as these are the parameters potentially biased when using two-stage methodologies. Besides, average computational times are also compared.

Broadly speaking, the conclusions for each dataset do not change due to the inferential methodology chosen. Point and interval estimates are relatively robust for the group parameter regardless of the estimation approach. The same pattern is also observed for the association parameter from the frequentist perspective, i.e., the three approaches seem to be equivalent. On the other hand, the Bayesian standard two-stage (BSTS) approach presents estimates that are a little different from the others, and which can potentially be

Table 7.1: Estimation and 95% confidence/credible interval for survival regression parameters and computational time from each inferential approach using three real datasets.

| Perspective | Approach | $\gamma_1$ (Group) | $\alpha$ (Association) | *ACT |
|---|---|---|---|---|
| prothro dataset | | | | |
| Frequentist | JS | -0.132 (-0.368, 0.104) | -2.292 (-2.333, -2.251) | 13.818 |
| | NTS | -0.164 (-0.404, 0.075) | -2.371 (-2.717, -2.024) | 1.511 |
| | STS | -0.153 (-0.392, 0.085) | -2.336 (-2.695, -1.976) | 0.013 |
| Bayesian | BJS | -0.159 (-0.408, 0.088) | -2.386 (-2.776, -1.967) | 21.499 |
| | NBTS1 | -0.170 (-0.453, 0.083) | -2.384 (-2.747, -2.018) | 9.632 |
| | NBTS2 | -0.168 (-0.418, 0.082) | -2.322 (-2.731, -1.914) | 13.463 |
| | BSTS | -0.134 (-0.370, 0.103) | -2.134 (-2.493, -1.775) | 4.862 |
| aids dataset | | | | |
| Frequentist | JS | 0.342 (0.054, 0.631) | -1.084 (-1.204, -0.964) | 3.712 |
| | NTS | 0.337 (0.042, 0.633) | -0.994 (-1.206, -0.782) | 1.391 |
| | STS | 0.322 (0.028, 0.615) | -0.938 (-1.148, -0.728) | 0.019 |
| Bayesian | BJS | 0.334 (0.026, 0.642) | -1.081 (-1.319, -0.843) | 11.031 |
| | NBTS1 | 0.335 (0.036, 0.634) | -0.988 (-1.201, -0.775) | 5.419 |
| | NBTS2 | 0.340 (0.025, 0.655) | -1.044 (-1.277, -0.811) | 5.224 |
| | BSTS | 0.333 (0.027, 0.639) | -0.970 (-1.176, -0.765) | 3.609 |
| pbc dataset | | | | |
| Frequentist | JS | 0.001 (-0.291, 0.292) | 1.198 (1.108, 1.288) | 1.793 |
| | NTS | 0.006 (-0.307, 0.319) | 1.219 (1.065, 1.374) | 0.472 |
| | STS | -0.014 (-0.325, 0.297) | 1.209 (1.055, 1.363) | 0.029 |
| Bayesian | BJS | -0.004 (-0.339, 0.331) | 1.226 (1.052, 1.402) | 8.504 |
| | NBTS1 | 0.019 (-0.284, 0.317) | 1.219 (1.071, 1.371) | 6.559 |
| | NBTS2 | -0.018 (-0.347, 0.311) | 1.230 (1.058, 1.403) | 4.844 |
| | BSTS | -0.000 (-0.317, 0.317) | 1.121 (0.976, 1.266) | 3.609 |

*ACT: Average Computational Time (in minutes).

seen as estimation biases. However, as expected, the standard two-stage methodologies, both frequentist and Bayesian, are the fastest, while our proposals (NTS, NBTS1, and NBTS2) take a little longer than them, but with quite reasonable times and less than those obtained with the joint specification approaches (JS and BJS).

In the context of each dataset, only AIDS treatments seem to be statistically different, in which the didanosine (ddI, experimental group) has a higher risk of death than the zalcitabine (ddC, control group). On the other hand, there is no doubt about the relevance of the association parameter in each context, indicating that analyzed longitudinal biomarkers effectively help explain the time to death.

# Chapter 8

# Simulation study

This chapter validates and compares our frequentist and Bayesian two-stage methodologies with their competitors using a broad simulation study. Scenarios are simulated using Algorithm 2 based on the joint model (6.1)-(6.2). For each scenario, we simulate 50 datasets with $n = 500$ (number of individuals), $m_{\min} = 5, 10, 20$ (minimum number of longitudinal observations), and $t_{\max} = 20$ (maximum observational time).

## 8.1 Scenarios

To complete the simulation scheme input arguments, we must specify the true parameter values, $\boldsymbol{\theta}_y = (\beta_0, \beta_1, \beta_2, \sigma^2)^\top$, $\boldsymbol{\theta}_b = \boldsymbol{\Sigma}$, and $\boldsymbol{\theta}_s = (\phi, \gamma_0, \gamma_1, \alpha)^\top$. To do this realistically, we use the estimated values of each parameter from the Bayesian joint specification (BJS) approach for the three datasets. Table 8.1 shows the true parameter values for each scenario.

Table 8.1: Setting true parameter values for each scenario.

| Parameter | Scenario I | Scenario II | Scenario III |
|---|---|---|---|
| $\beta_0$ | 4.276 | 2.473 | 0.550 |
| $\beta_1$ | -0.004 | -0.039 | 0.190 |
| $\beta_2$ | -0.097 | 0.085 | -0.120 |
| $\sigma^2$ | 0.069 | 0.135 | 0.121 |
| $\Sigma_{11}$ | 0.093 | 0.757 | 0.997 |
| $\Sigma_{22}$ | 0.004 | 0.001 | 0.033 |
| $\Sigma_{12}$ | -0.001 | 0.002 | 0.435 |
| $\phi$ | 0.937 | 1.267 | 1.084 |
| $\gamma_0$ | 8.500 | -2.204 | -4.302 |
| $\gamma_1$ | -0.159 | 0.334 | -0.004 |
| $\alpha$ | -2.372 | -1.081 | 1.226 |

Note that Scenarios I, II and III are based on the joint modeling for the prothro, aids and pbc datasets, respectively, used in Chapter 7.

## 8.2 Results

Similar to Chapter 7, SIMEX parameters are defined as $K = 50$ and $\Delta = [0.5, 1, 1.5, 2]$. In addition, NTS and NBTS1 user-specified parameters, $p$ and $\eta$, are chosen for each scenario from a sensitivity analysis presented in Appendices A.1 and B.1. Specifically, $p = 0.9$ and $\eta = 10$ for Scenario I, $p = 0.9$ and $\eta = 20$ for Scenario II, and $p = 0.1$ and $\eta = 30$ for Scenario III. For Bayesian models, the MCMC configuration is set as in Chapter 7. Table 8.2 shows the comparative results for $\gamma_1$ and $\alpha$ using Scenario I.

Table 8.2: Scenario I: Estimation and 95% confidence/credible interval for survival regression parameters and computational time from each inferential approach.

| $m_{min}$ | Approach | $\gamma_1$ (Group) True Value: -0.159 | $\alpha$ (Association) True Value: -2.372 | *ACT |
|---|---|---|---|---|
| Frequentist Perspective | | | | |
| 5 | JS | -0.144 (-0.364, 0.075) | -2.245 (-2.283, -2.207) | 7.584 |
| | NTS | -0.150 (-0.515, 0.215) | -2.330 (-2.551, -2.108) | 2.188 |
| | STS | -0.142 (-0.363, 0.079) | -2.216 (-2.580, -1.851) | 0.143 |
| 10 | JS | -0.162 (-0.382, 0.058) | -2.325 (-2.363, -2.287) | 10.125 |
| | NTS | -0.170 (-0.527, 0.188) | -2.405 (-2.628, -2.181) | 2.553 |
| | STS | -0.162 (-0.385, 0.061) | -2.320 (-2.677, -1.964) | 0.192 |
| 20 | JS | -0.168 (-0.385, 0.048) | -2.276 (-2.313, -2.238) | 13.389 |
| | NTS | -0.174 (-0.521, 0.173) | -2.343 (-2.565, -2.121) | 3.461 |
| | STS | -0.170 (-0.391, 0.052) | -2.291 (-2.639, -1.944) | 0.262 |
| Bayesian Perspective | | | | |
| 5 | BJS | -0.155 (-0.386, 0.076) | -2.471 (-2.861, -2.081) | 18.030 |
| | NBTS1 | -0.149 (-0.383, 0.084) | -2.377 (-2.750, -2.004) | 9.234 |
| | NBTS2 | -0.148 (-0.381, 0.086) | -2.359 (-2.745, -1.973) | 13.670 |
| | BSTS | -0.142 (-0.369, 0.084) | -2.272 (-2.631, -1.914) | 6.552 |
| 10 | BJS | -0.155 (-0.386, 0.076) | -2.429 (-2.804, -2.054) | 21.660 |
| | NBTS1 | -0.154 (-0.388, 0.080) | -2.411 (-2.777, -2.046) | 10.412 |
| | NBTS2 | -0.151 (-0.380, 0.078) | -2.377 (-2.742, -2.011) | 16.240 |
| | BSTS | -0.149 (-0.375, 0.078) | -2.313 (-2.660, -1.996) | 8.416 |
| 20 | BJS | -0.154 (-0.381, 0.072) | -2.439 (-2.799, -2.078) | 27.520 |
| | NBTS1 | -0.155 (-0.390, 0.081) | -2.464 (-2.833, -2.094) | 13.700 |
| | NBTS2 | -0.150 (-0.377, 0.076) | -2.399 (-2.756, -2.042) | 18.880 |
| | BSTS | -0.148 (-0.372, 0.076) | -2.354 (-2.703, -2.006) | 11.290 |

*ACT: Average Computational Time (in minutes).

Similar to real applications (see Chapter 7), the group parameter estimation differences are negligible and very close to the true value of $\gamma_1$, showing once again that this parameter is not significantly affected when using two-stage methods. Increasing the minimum number of longitudinal observations ($m_{\min}$) does not alter the quality of the $\gamma_1$ estimation either. Appendices A.2, B.2 and C.1 also study this behavior through a bias analysis using point estimates of each approach by simulated dataset. Visibly, frequentist approaches (JS, NTS and STS) with $m_{\min} = 5$ are the only ones that show some estimation bias, but this is quickly corrected by increasing $m_{\min}$ to 10 or 20.

Table 8.2 also shows that frequentist point estimates of the association parameter are relatively similar. However, the NTS confidence intervals are much narrower than those from the STS approach. On the other hand, from the Bayesian approach, BSTS point estimates are notoriously biased, both in comparison to the true parameter value and to the value estimated by the Bayesian joint specification (BJS) approach. Our Bayesian two-stage proposals reduce the estimation bias, having NBTS1 point estimates closer to those of BJS. We can also appreciate that as $m_{\min}$ increases, the NBTS2 point estimate resembles that of BJS, as demonstrated in Proposition 4.4.1.

As for the processing time, standard two-stage (STS and BSTS) approaches are the fastest and joint specifications (JS and BJS) are the most time-consuming. In the frequentist perspective, as expected, our proposal is a little more computationally costly than the STS, but much faster than the JS approach. From a Bayesian point of view, NBTS1 is faster than the NBTS2 approach and only slightly slower than the BSTS. Still, both Bayesian proposals drastically reduce the time consumed by the BJS approach.

Table 8.3 presents the inferential comparisons for $\gamma_1$ and $\alpha$ using Scenario II as well as the average computational times of each estimation approach.

Table 8.3: Scenario II: Estimation and 95% confidence/credible interval for survival regression parameters and computational time from each inferential approach.

| $m_{min}$ | Approach | $\gamma_1$ (Group) True Value: 0.334 | $\alpha$ (Association) True Value: -1.081 | *ACT |
|---|---|---|---|---|
| Frequentist Perspective | | | | |
| 5 | JS | 0.344 (-0.113, 0.802) | -1.113 (-1.261, -0.984) | 9.251 |
| | NTS | 0.346 (-0.147, 0.645) | -1.133 (-1.637, -0.630) | 3.744 |
| | STS | 0.346 (-0.158, 0.849) | -1.133 (-1.432, -0.834) | 0.177 |
| 10 | JS | 0.343 (-0.113, 0.799) | -1.110 (-1.258, -0.962) | 11.489 |
| | NTS | 0.343 (-0.145, 0.641) | -1.130 (-1.634, -0.626) | 4.368 |
| | STS | 0.343 (-0.161, 0.846) | -1.130 (-1.428, -0.832) | 0.256 |
| 20 | JS | 0.344 (-0.112, 0.800) | -1.106 (-1.254, -0.958) | 15.252 |
| | NTS | 0.345 (-0.148, 0.642) | -1.121 (-1.625, -0.617) | 5.226 |
| | STS | 0.345 (-0.159, 0.849) | -1.121 (-1.417, -0.824) | 0.285 |
| Bayesian Perspective | | | | |
| 5 | BJS | 0.340 (-0.178, 0.859) | -1.131 (-1.432, -0.831) | 18.170 |
| | NBTS1 | 0.352 (-0.151, 0.855) | -1.127 (-1.422, -0.832) | 10.854 |
| | NBTS2 | 0.342 (-0.159, 0.863) | -1.130 (-1.429, -0.831) | 11.485 |
| | BSTS | 0.351 (-0.167, 0.869) | -1.120 (-1.416, -0.824) | 9.029 |
| 10 | BJS | 0.343 (-0.173, 0.859) | -1.133 (-1.429, -0.836) | 23.610 |
| | NBTS1 | 0.348 (-0.164, 0.861) | -1.131 (-1.428, -0.834) | 14.290 |
| | NBTS2 | 0.350 (-0.171, 0.872) | -1.131 (-1.429, -0.834) | 15.040 |
| | BSTS | 0.345 (-0.167, 0.856) | -1.122 (-1.415, -0.829) | 12.222 |
| 20 | BJS | 0.342 (-0.173, 0.856) | -1.130 (-1.422, -0.830) | 31.500 |
| | NBTS1 | 0.353 (-0.158,0.864) | -1.127 (-1.419, -0.835) | 19.620 |
| | NBTS2 | 0.342 (-0.165, 0.869) | -1.130 (-1.413, -0.824) | 21.320 |
| | BSTS | 0.342 (-0.173, 0.856) | -1.123 (-1.418, -0.827) | 17.810 |

*ACT: Average Computational Time (in minutes).

Again, the group parameter point estimate is extremely robust from all inferential methodologies, with very little difference when using our first Bayesian proposal. It is also worth noting that the 95% confidence/credible intervals for $\gamma_1$ are quite stable, even when increasing the minimum number of longitudinal observations ($m_{\min}$). The same happens with point and interval estimates for the association parameter from the frequentist approaches. This pattern can also be seen in the estimation bias analysis presented in Appendix A.2. In this case, the results of the NTS and STS estimations are equivalent and, when comparing the computational times, it would be preferable to use the (frequentist) standard approach.

Focusing on association parameter estimations from the Bayesian paradigm, the differences are subtle, except that NBTS1 is not able to estimate as accurately as other approaches. NBTS1's regular performance can also be seen in Figure B.11 of Appendix B.2. Additionally, NBTS2 produces point and interval estimates very similar to the BJS approach, even with $m_{\min} = 5$.

Table 8.3 also shows the average computational time. Again, the inferential process is performed faster by the standard two-stage approaches and slower by the joint specifications approaches. Since NBTS1 does not estimate the survival regression parameters very well, then the most interesting processing time comparisons are between the BJS, BSTS and NBTS2 methods. Specifically, our second Bayesian proposal spends around 30% more time than its two-stage competitor (BSTS) and 70% less than the joint specification (BJS) approach.

Finally, Table 8.4 compares point and interval estimates for survival regression parameters (group and association) and computational time from each inferential approach using Scenario III.

Table 8.4: Scenario III: Estimation and 95% confidence/credible interval for survival regression parameters and computational time from each inferential approach.

| $m_{min}$ | Approach | $\gamma_1$ (Group) True Value: -0.004 | $\alpha$ (Association) True Value: 1.226 | *ACT |
|---|---|---|---|---|
| Frequentist Perspective | | | | |
| 5 | JS | -0.020 (-0.299, 0.303) | 1.205 (1.086, 1.324) | 8.596 |
| | NTS | -0.016 (-0.190, 0.158) | 1.254 (0.909, 1.598) | 2.919 |
| | STS | -0.016 (-0.361, 0.329) | 1.253 (1.079, 1.428) | 0.185 |
| 10 | JS | -0.021 (-0.324, 0.281) | 1.201 (1.083, 1.320) | 12.881 |
| | NTS | -0.026 (-0.200, 0.148) | 1.261 (0.917, 1.606) | 4.827 |
| | STS | -0.026 (-0.371, 0.318) | 1.261 (1.087, 1.435) | 0.395 |
| 20 | JS | -0.009 (-0.308, 0.290) | 1.193 (1.077, 1.310) | 16.617 |
| | NTS | -0.023 (-0.195, 0.149) | 1.258 (0.913, 1.602) | 5.226 |
| | STS | -0.023 (-0.368, 0.322) | 1.257 (1.085, 1.430) | 0.508 |
| Bayesian Perspective | | | | |
| 5 | BJS | -0.017 (-0.368, 0.335) | 1.245 (1.070, 1.420) | 22.230 |
| | NBTS1 | -0.009 (-0.355, 0.336) | 1.236 (1.065, 1.407) | 11.020 |
| | NBTS2 | -0.015 (-0.359, 0.336) | 1.242 (1.066, 1.418) | 12.880 |
| | BSTS | -0.007 (-0.355, 0.341) | 1.223 (1.052, 1.393) | 9.797 |
| 10 | BJS | -0.027 (-0.378, 0.324) | 1.258 (1.083, 1.433) | 30.185 |
| | NBTS1 | -0.019 (-0.366, 0.328) | 1.256 (1.083, 1.428) | 13.640 |
| | NBTS2 | -0.025 (-0.376, 0.325) | 1.255 (1.078, 1.431) | 16.010 |
| | BSTS | -0.021 (-0.374, 0.332) | 1.241 (1.068, 1.415) | 12.540 |
| 20 | BJS | -0.024 (-0.375, 0.327) | 1.255 (1.080, 1.430) | 33.200 |
| | NBTS1 | -0.019 (-0.365, 0.327) | 1.261 (1.089, 1.433) | 21.920 |
| | NBTS2 | -0.023 (-0.376, 0.332) | 1.254 (1.079, 1.426) | 23.710 |
| | BSTS | -0.023 (-0.373, 0.327) | 1.242 (1.073, 1.418) | 19.310 |

*ACT: Average Computational Time (in minutes).

The frequentist results in Table 8.4 are similar to those in Scenario II, where both two-stage strategies can be considered equivalent and very similar to the joint specification approach. For Bayesian models, the group parameter is still robustly estimated, with a slight difference in point estimation using the NBTS1 approach. On the other hand, BJS, NBTS1 and NBTS2 estimate the association parameter practically equal, while the BSTS presents a small bias in comparisons with the other approaches. In this case, one of the novel two-stage proposals would be preferable, as their computational times are much lower than the Bayesian joint specification approach.

In summary, our frequentist and Bayesian proposals are better than or equal to the standard two-stage methods and their computational times are always much lower than the joint specification approaches.

# Chapter 9

# Conclusions and discussion

In this thesis, we have developed three novel two-stage approaches for joint models of longitudinal and survival outcomes in order to reduce or eliminate the estimation bias, quite common in standard two-stage approaches, with an acceptable computational time.

First, from a frequentist perspective, we have introduced a novel two-stage (NTS) approach based on the simulation-extrapolation (SIMEX) algorithm. Our approach was shown to be promising, since it is more accurate than the standard two-stage (STS) approach and the processing time is much less than the joint specification (JS) approach. In addition, NTS is easily implemented using standard R-packages for each submodel (longitudinal and survival). In our simulation study, we have found that the inference for the survival submodel group parameter ($\gamma_1$) is robustly estimated regardless of the estimation approach. This result is expected since this parameter does not depend on shared information and therefore it is not affected by ignoring the joint nature between both processes. On the other hand, the association parameter ($\alpha$) is quite sensitive, sometimes producing

sharply biased estimates with both the STS approach and the JS approach (see the results of Scenario I in Chapter 8). Still, even in these most extreme cases, our methodology has satisfactorily reduced the estimation bias. However, the specification of the SIMEX measurement error variance, through the parameter $p$, by the user is the critical drawback of our approach. We partially circumvent this problem by proposing an ad-hoc estimation mechanism for this parameter. Even so, this is still an open issue to be improved.

From a Bayesian perspective, our first two-stage proposal (NBTS1) has also proved to be quite adequate. Specifically, point and interval estimations are very similar to the Bayesian joint specification (BJS) approach, but the computational time is significantly less than that of the BJS. Again, $\gamma_1$ has been well estimated with all methodologies, while $\alpha$ has been somewhat biased when estimated from the Bayesian standard two-stage (BSTS) approach. The specification of the informative prior variance ($\eta$) for the fixed-effects may be critical drawback of our Bayesian approach. In our simulation study, the set value has produced quite satisfactory results. However, we would like to reinforce to the reader that this choice depends on the scale of the problem, in which the value used in this thesis can cause large perturbations in other applications and, therefore, a smaller value should be preferred.

Our second Bayesian proposal (NBST2) uses the estimations from the longitudinal submodel to specify an informative prior distribution for the random effects when estimating them within the survival submodel. In addition, as a bias correction mechanism, NBST2 incorporates the longitudinal likelihood function in the second stage, where its fixed effects are set according to the estimation using only the longitudinal submodel. Note that this approach does not require the user to set any parameters, such as $p$ in NTS and $\eta$ in NBTS1. Moreover, we have shown in Proposition 4.4.1 that this second Bayesian

approach resembles a joint inference as the number of repeated longitudinal measurements per individual. Furthermore, NBTS2 was shown to be successful, producing satisfactory results both in real applications and in simulated scenarios.

Overall, the group parameter has not been significantly affected by the use of two-stage approaches. However, we have once again observed that the association parameter is biased when the standard two-stage approaches (STS and BSTS) are used. This issue can be solved by replacing STS/BSTS with NTS/NBTS1/NBTS2. In all scenarios, our proposals have been much faster than the joint specification approaches (JS and BJS).

In future directions, it would be interesting to apply our proposals in more complex longitudinal (e.g., skewed or multiple longitudinal data) and survival (e.g., competing-risks or multistate data) submodels than those employed here. Hence, we would be able to try determining the limits of the methodology. A further topic for research would be to explore other frequentist and Bayesian methods to increase the speed of the computation involved in fitting each submodel.

# Appendix A

# Simulation study: NTS approach

## A.1 Sensitivity analysis for $p$

### Scenario I

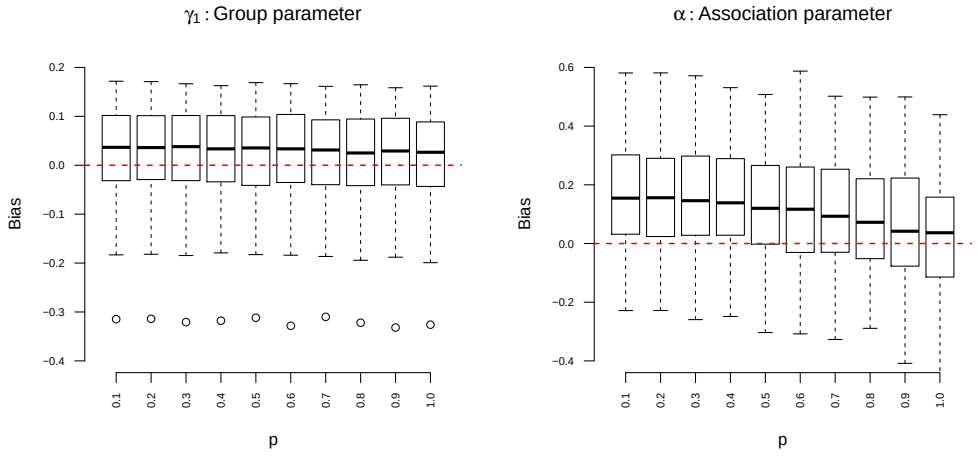


Figure A.1: Scenario I with $m_{\min} = 5$: Bias for survival regression parameter estimates considering different values of $p$. The dashed horizontal line indicates no bias.

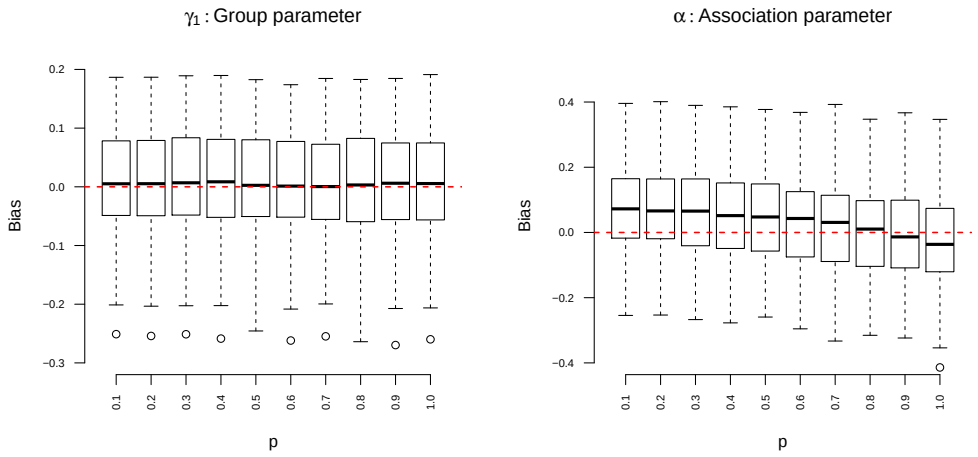


Figure A.2: Scenario I with $m_{\min} = 10$: Bias for survival regression parameter estimates considering different values of $p$. The dashed horizontal line indicates no bias.

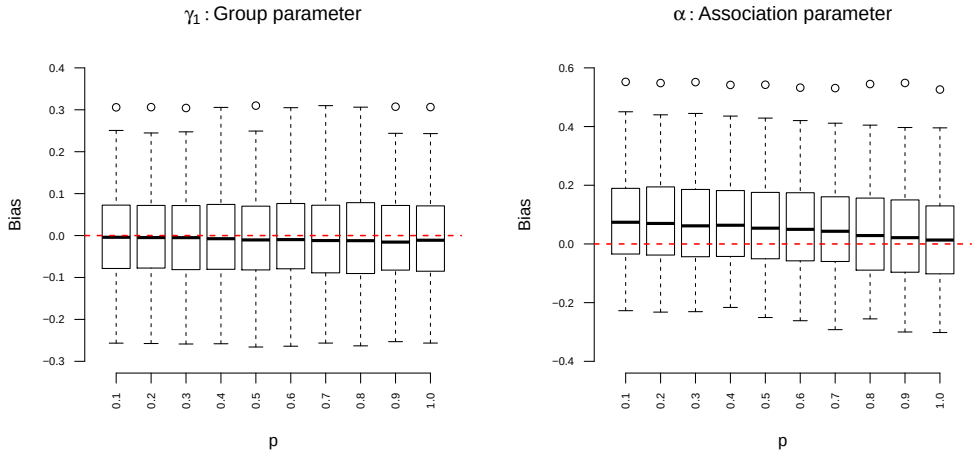


Figure A.3: Scenario I with $m_{\min} = 20$: Bias for survival regression parameter estimates considering different values of $p$. The dashed horizontal line indicates no bias.

## Scenario II

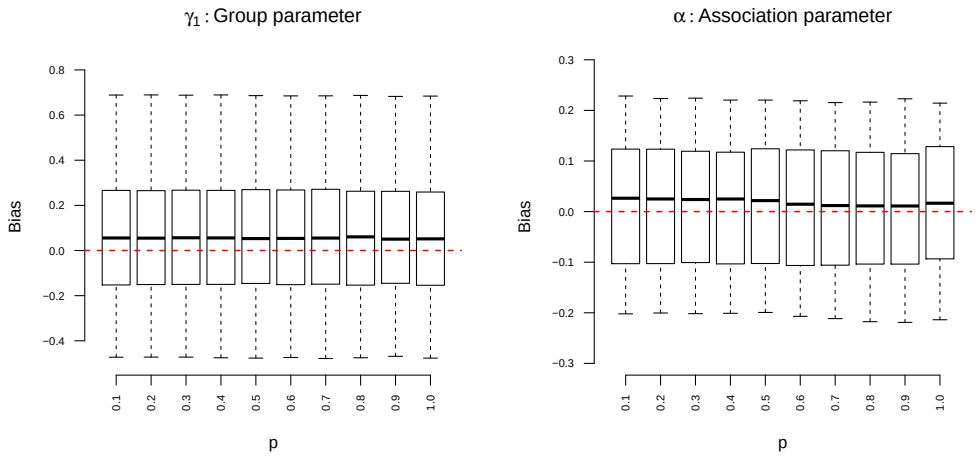


Figure A.4: Scenario II with $m_{\min} = 5$: Bias for survival regression parameter estimates considering different values of $p$. The dashed horizontal line indicates no bias.

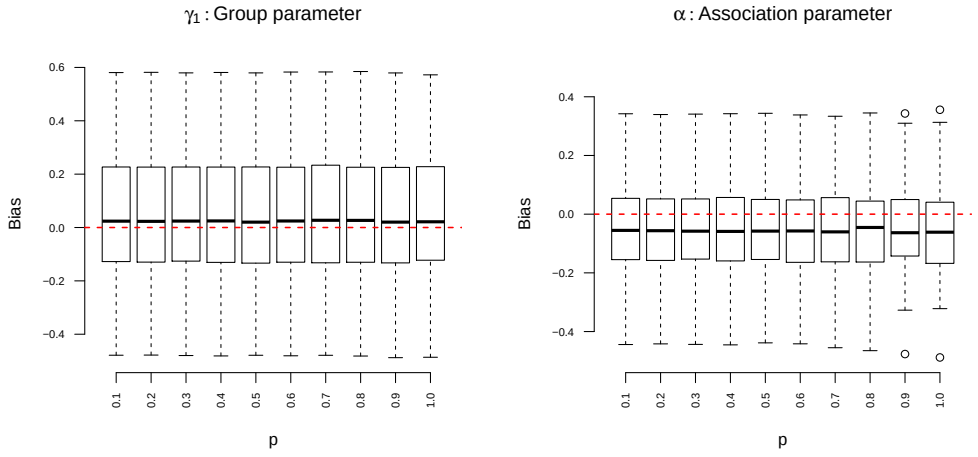


Figure A.5: Scenario II with $m_{\min} = 10$: Bias for survival regression parameter estimates considering different values of $p$. The dashed horizontal line indicates no bias.

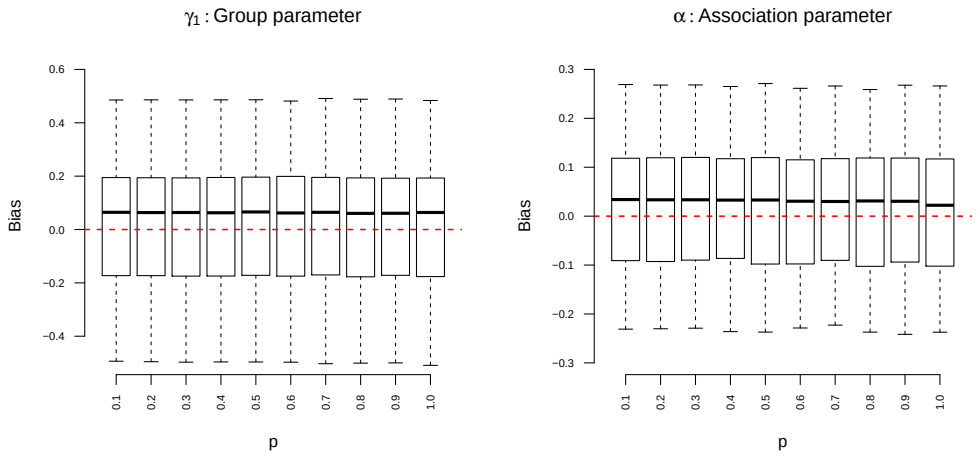


Figure A.6: Scenario II with $m_{\min} = 20$: Bias for survival regression parameter estimates considering different values of $p$. The dashed horizontal line indicates no bias.

## Scenario III

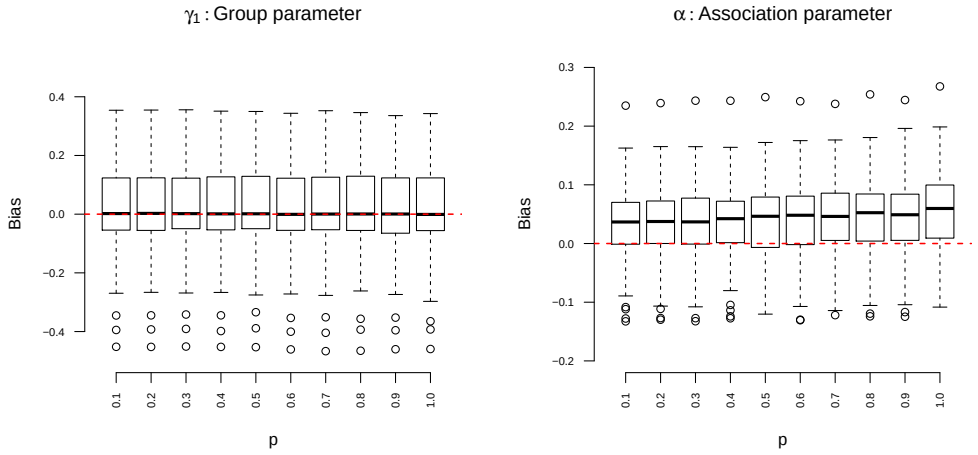


Figure A.7: Scenario III with $m_{\min} = 5$: Bias for survival regression parameter estimates considering different values of $p$. The dashed horizontal line indicates no bias.

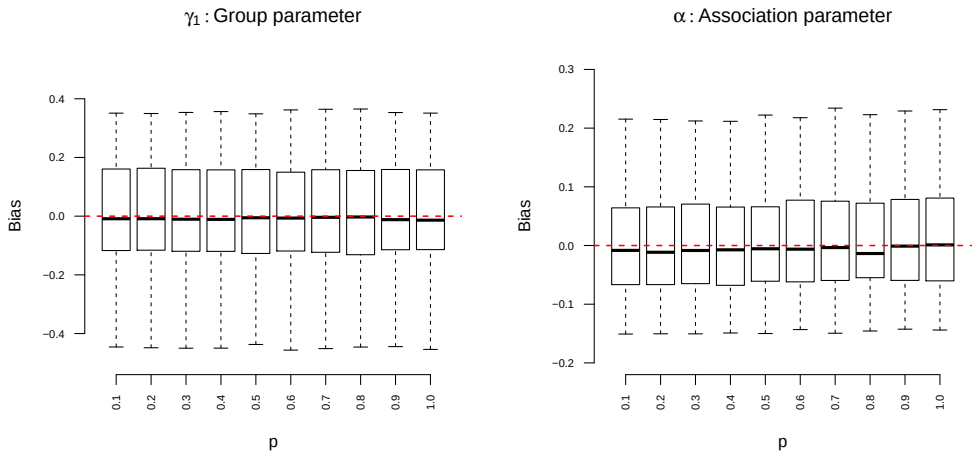


Figure A.8: Scenario III with $m_{\min} = 10$: Bias for survival regression parameter estimates considering different values of $p$. The dashed horizontal line indicates no bias.

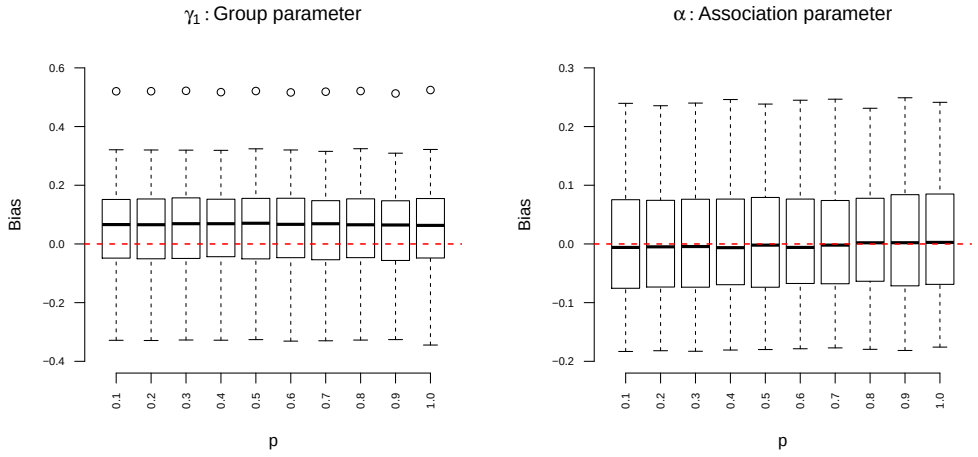


Figure A.9: Scenario III with $m_{\min} = 20$: Bias for survival regression parameter estimates considering different values of $p$. The dashed horizontal line indicates no bias.

## A.2 Estimation bias analysis by simulation scenario

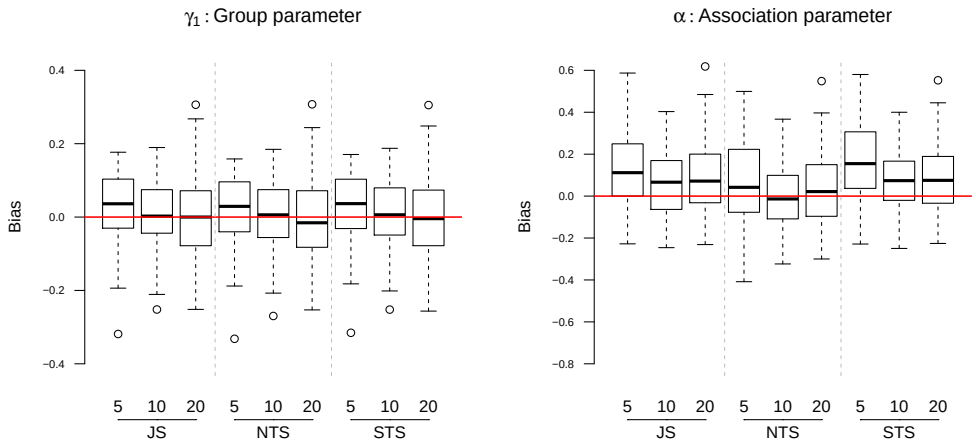


Figure A.10: Scenario I with $p = 0.9$: Bias for survival regression parameter estimates from JS, NTS and STS approaches. The dashed horizontal line indicates no bias.

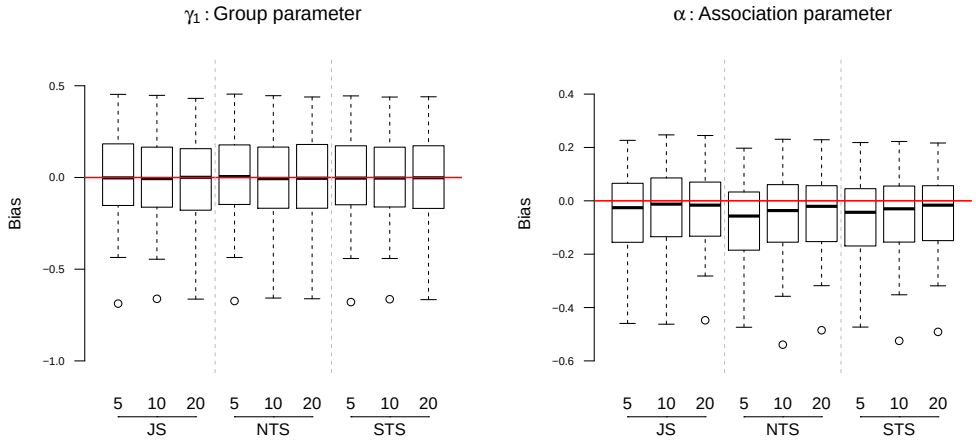


Figure A.11: Scenario II with $p = 0.9$: Bias for survival regression parameter estimates from JS, NTS and STS approaches. The dashed horizontal line indicates no bias.

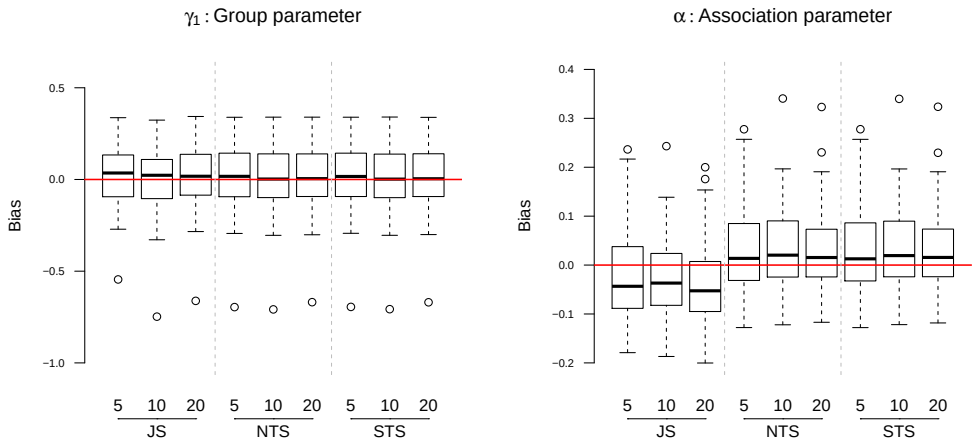


Figure A.12: Scenario III with $p = 0.1$: Bias for survival regression parameter estimates from JS, NTS and STS approaches. The dashed horizontal line indicates no bias.

# Appendix B

# Simulation study: NBTS1 approach

## B.1 Sensitivity analysis for $\eta$

### Scenario I

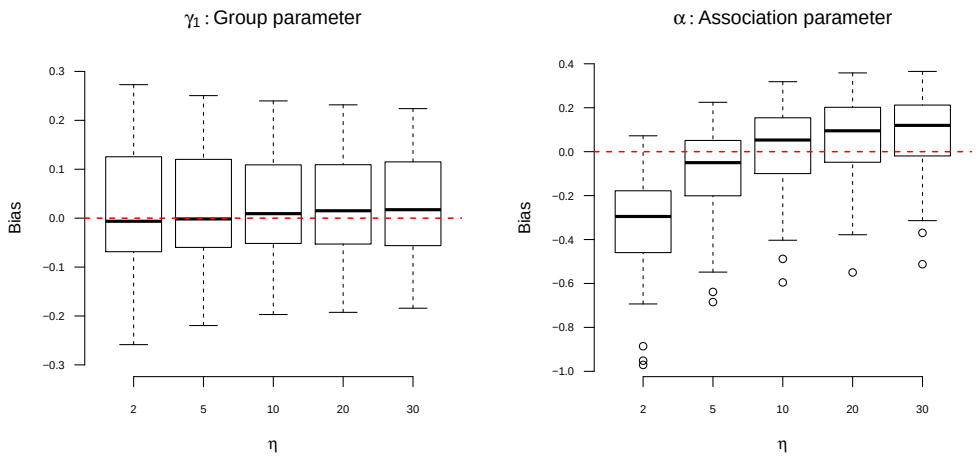


Figure B.1: Scenario I with $m_{\min} = 5$: Bias for survival regression parameter estimates considering different values of $\eta$. The dashed horizontal line indicates no bias.

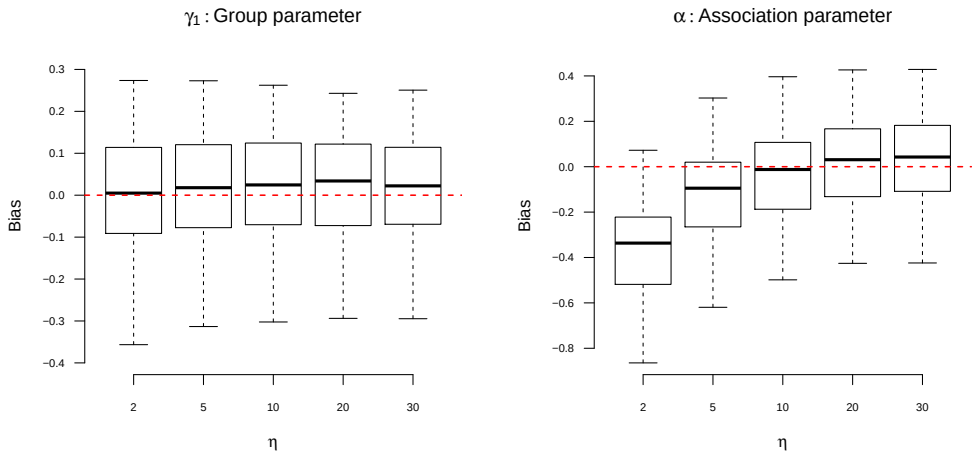


Figure B.2: Scenario I with $m_{\min} = 10$: Bias for survival regression parameter estimates considering different values of $\eta$. The dashed horizontal line indicates no bias.

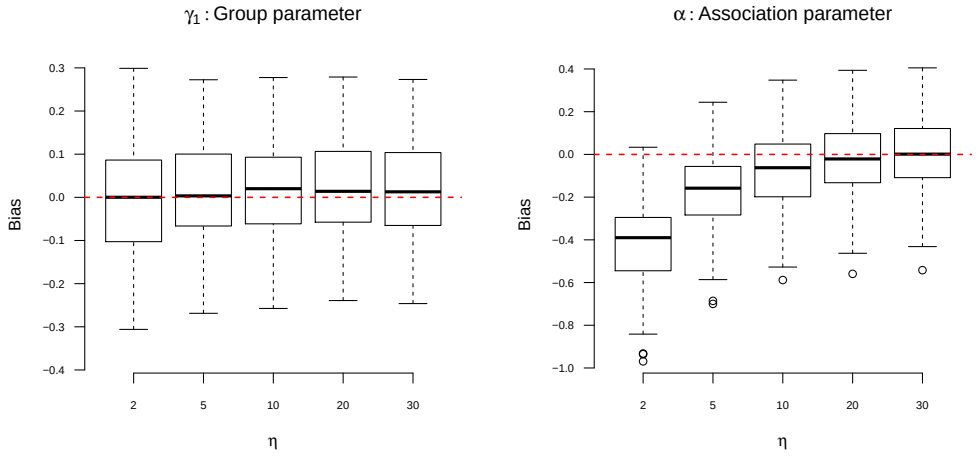


Figure B.3: Scenario I with $m_{\min} = 20$: Bias for survival regression parameter estimates considering different values of $\eta$. The dashed horizontal line indicates no bias.

## Scenario II

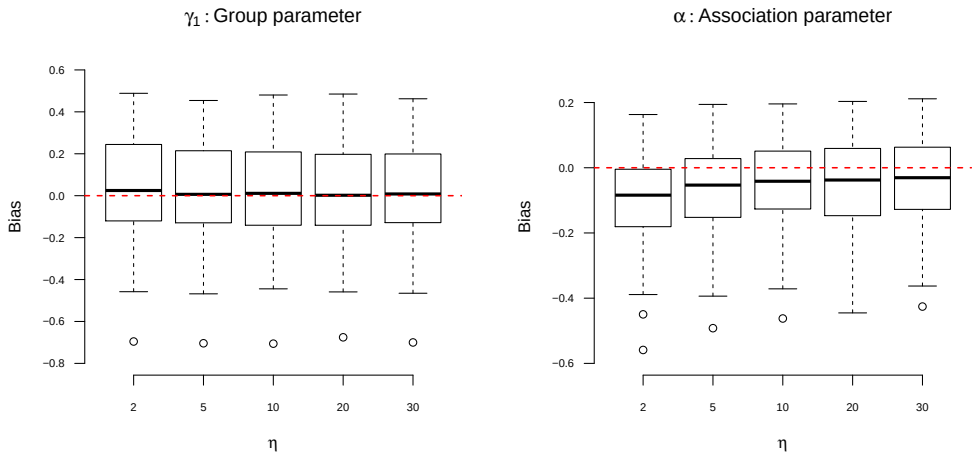


Figure B.4: Scenario II with $m_{\min} = 5$: Bias for survival regression parameter estimates considering different values of $\eta$. The dashed horizontal line indicates no bias.

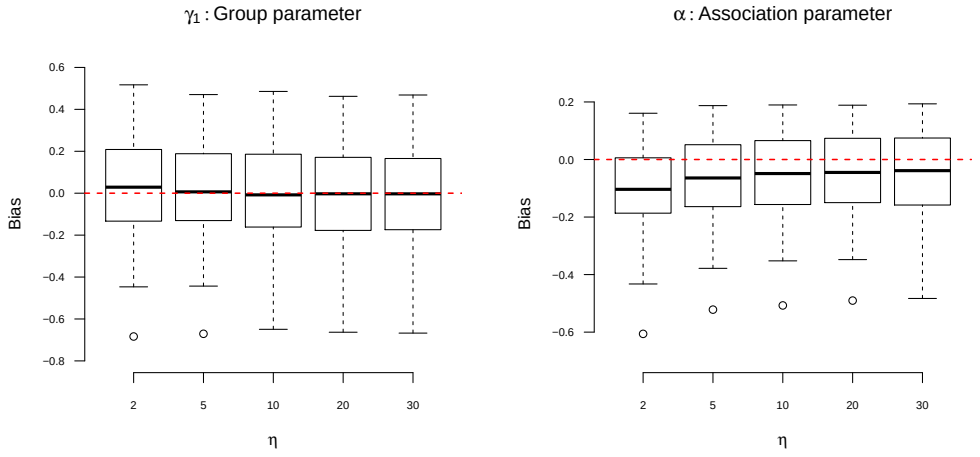


Figure B.5: Scenario II with $m_{\min} = 10$: Bias for survival regression parameter estimates considering different values of $\eta$. The dashed horizontal line indicates no bias.

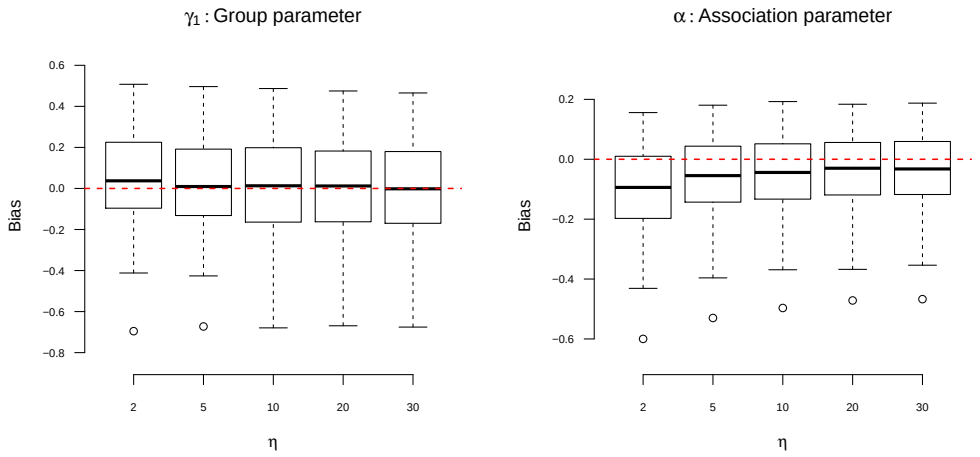


Figure B.6: Scenario II with $m_{\min} = 20$: Bias for survival regression parameter estimates considering different values of $\eta$. The dashed horizontal line indicates no bias.

## Scenario III

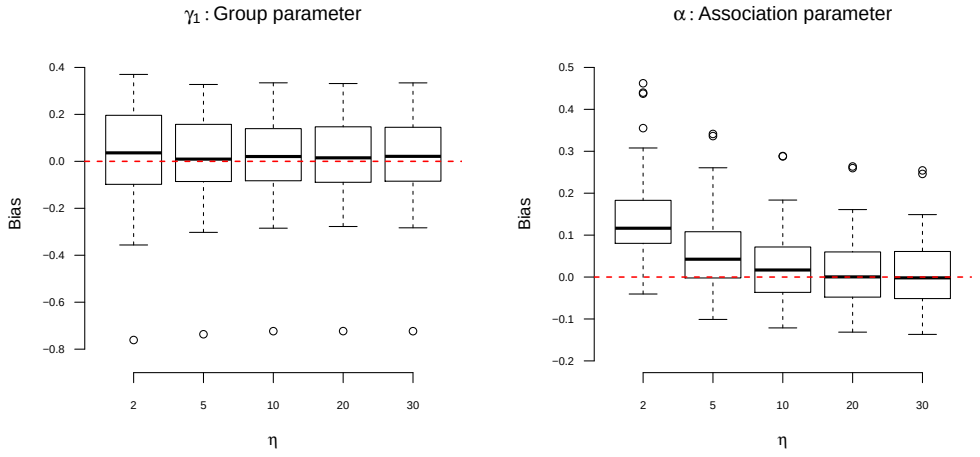


Figure B.7: Scenario III with $m_{\min} = 5$: Bias for survival regression parameter estimates considering different values of $\eta$. The dashed horizontal line indicates no bias.

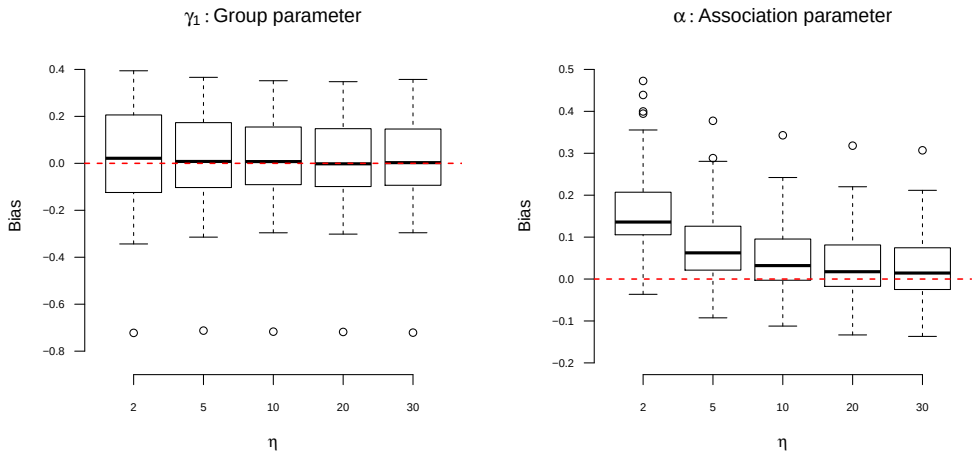


Figure B.8: Scenario III with $m_{\min} = 10$: Bias for survival regression parameter estimates considering different values of $\eta$. The dashed horizontal line indicates no bias.

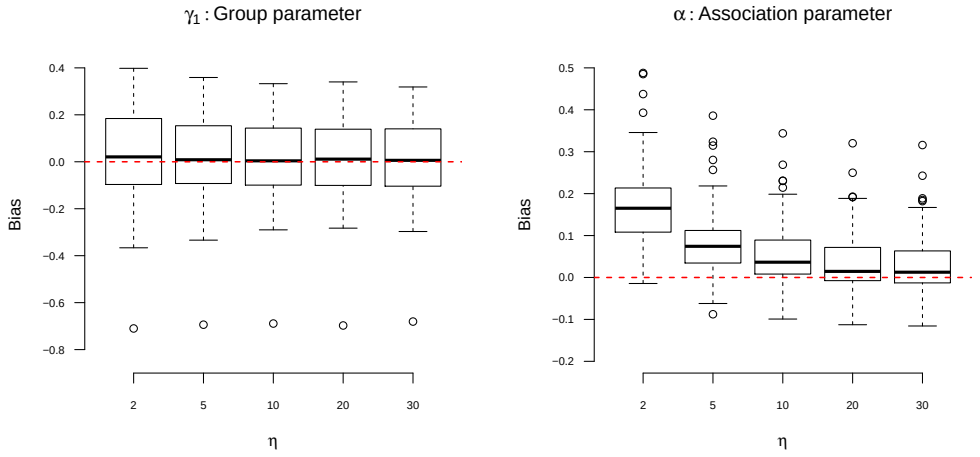


Figure B.9: Scenario III with $m_{\min} = 20$: Bias for survival regression parameter estimates considering different values of $\eta$. The dashed horizontal line indicates no bias.

## B.2 Estimation bias analysis by simulation scenario

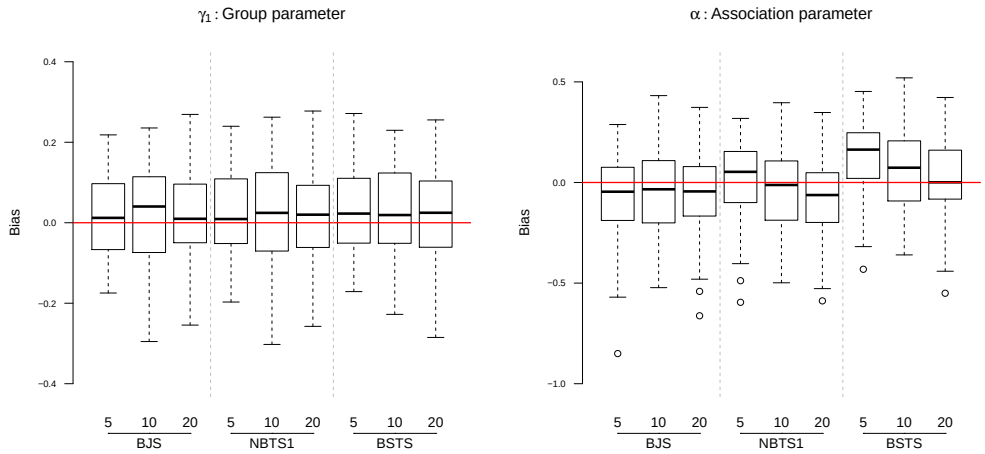


Figure B.10: Scenario I with $\eta = 10$: Bias for survival regression parameter estimates from BJS, NBTS1 and BSTS approaches. The dashed horizontal line indicates no bias.

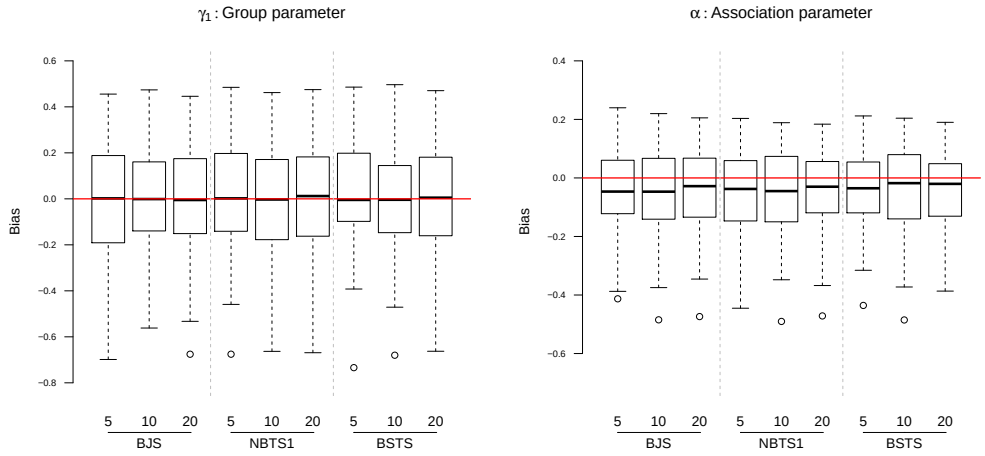


Figure B.11: Scenario II with $\eta = 20$: Bias for survival regression parameter estimates from BJS, NBTS1 and BSTS approaches. The dashed horizontal line indicates no bias.

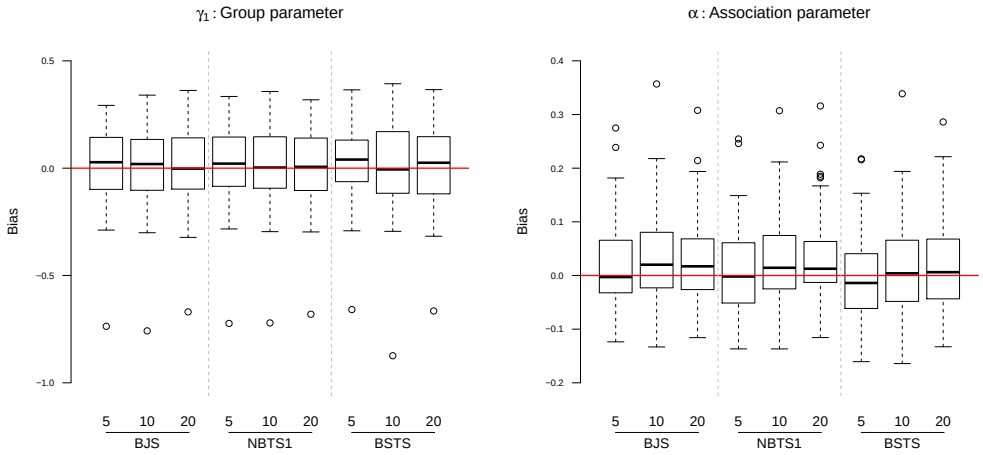


Figure B.12: Scenario III with $\eta = 30$: Bias for survival regression parameter estimates from BJS, NBTS1 and BSTS approaches. The dashed horizontal line indicates no bias.

# Appendix C

# Simulation study: NBTS2 approach

## C.1 Estimation bias analysis by simulation scenario

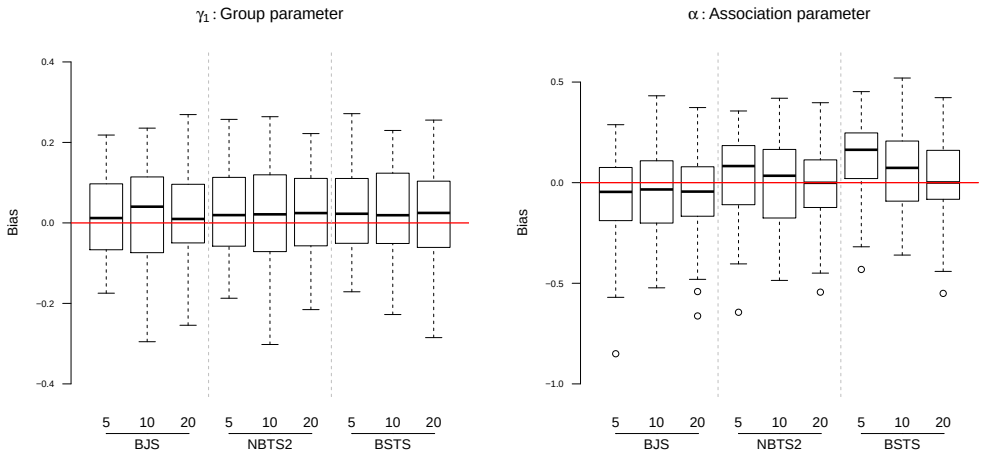


Figure C.1: Scenario I: Bias for survival regression parameter estimates from BJS, NBTS2 and BSTS approaches. The dashed horizontal line indicates no bias.

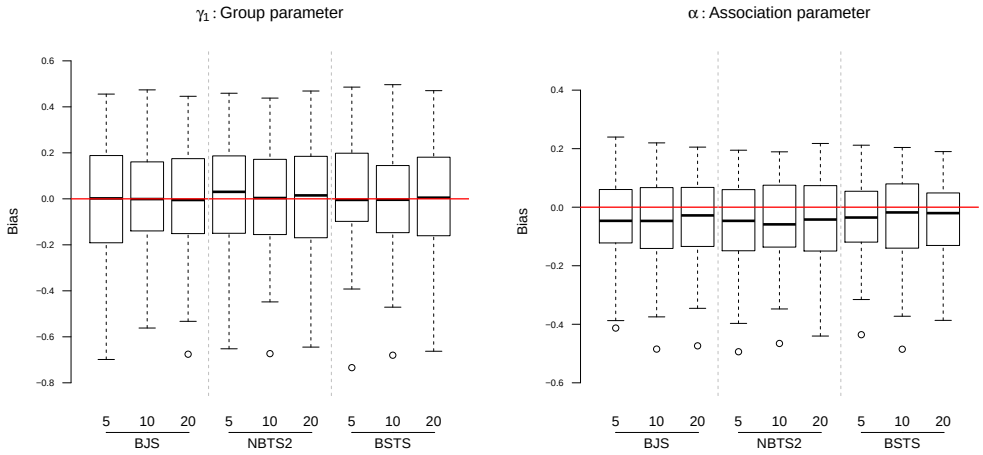


Figure C.2: Scenario II: Bias for survival regression parameter estimates from BJS, NBTS2 and BSTS approaches. The dashed horizontal line indicates no bias.

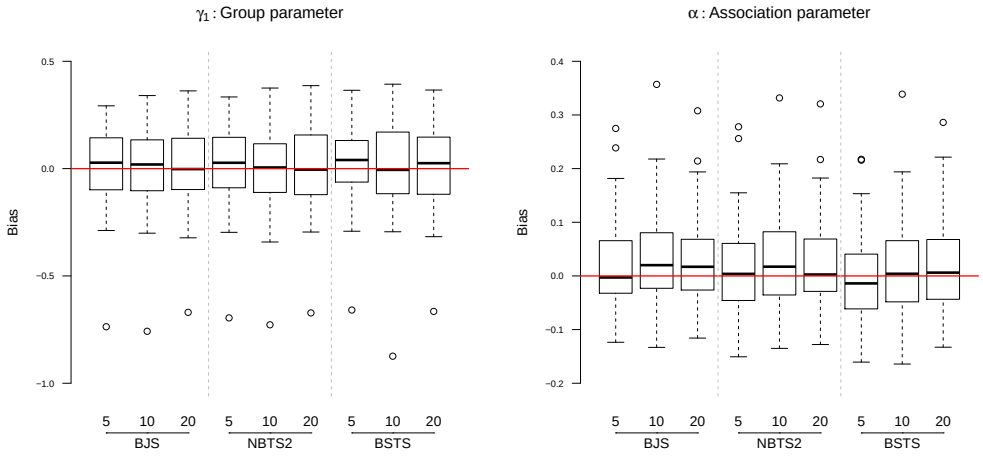


Figure C.3: Scenario III: Bias for survival regression parameter estimates from BJS, NBTS2 and BSTS approaches. The dashed horizontal line indicates no bias.

# Bibliography

Albert, P. and Shih, J. (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics*, 66(3):983–987.

Alsefri, M., Sudell, M., García-Fiñana, M., and Kolamunnage-Dona, R. (2020). Bayesian joint modelling of longitudinal and time to event data: a methodological review. *BMC Medical Research Methodology*, 20(94):1–17.

Alvares, D. (2017). *Sequential Monte Carlo methods in Bayesian joint models for longitudinal and time-to-event data*. PhD thesis, University of Valencia.

Alvares, D., Armero, C., Forte, A., and Chopin, N. (2021). Sequential Monte Carlo methods in Bayesian joint models for longitudinal and time-to-event data. *Statistical Modelling*, 21(1–2):161–181.

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer, 1st edition.

Andrinopoulou, E., Eilers, P., Takkenberg, J., and Rizopoulos, D. (2018). Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines. *Biometrics*, 74(2):685–693.

Apanasovich, T., Carroll, R., and Maity, A. (2009). SIMEX and standard error estimation in semiparametric measurement error models. *Electronic Journal of Statistics*, 3:318–348.

Armero, C., Forte, A., Perpiñán, H., Sanahuja, M., and Agustí, S. (2018). Bayesian joint modeling for assessing the progression of chronic kidney disease in children. *Statistical Methods in Medical Research*, 27(1):298–311.

Barrett, J., Diggle, P., Henderson, R., and Taylor-Robinson, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 77(1):131–148.

Carroll, R. and Küchenhoff, H. (1995). Approximative methods for regression models with errors in the covariates. In *XploRe: an interactive statistical computing environment*, pages 275–285. Springer.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman & Hall/CRC, 2nd edition.

Cook, J. and Stefanski, L. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cox, D. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Crowther, M. and Lambert, P. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32(23):4118–4134.

Donnelly, C., McFetridge, L., Marshall, A., and Mitchell, H.J. (2018). A two-stage approach to the joint analysis of longitudinal and survival data utilising the Coxian phase-type distribution. *Statistical Methods in Medical Research*, 27(12):3577–3594.

Elashoff, R., Li, G., and Li, N. (2016). *Joint modeling of longitudinal and time-to-event data*. Chapman & Hall/CRC, 1st edition.

Faucett, C. and Thomas, D. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15(15):1663–1685.

Furgal, K., Sen, A., and Taylor, J. (2019). Review and comparison of computational approaches for joint longitudinal and time-to-event models. *International Statistical Review*, 87(2):393–418.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian data analysis*. Chapman & Hall, 3rd edition.

Goldman, A., Carlin, B., Crane, L., Launer, C., Korvick, J., Deyton, L., and Abrams, D. (1996). Response of CD4 lymphocytes and clinical consequences of treatment using ddI or ddC in patients with advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11(2):161–169.

Gould, A., Boye, M., Crowther, M., Ibrahim, J., Quartey, G., Micallef, S., and Bois, F. (2015). 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.

Guler, I., Faes, C., Cadarso-Suárez, C., Teixeira, L., Rodrigues, A., and Mendonça, D. (2017). Two-stage model for multivariate longitudinal and survival data with application to nephrology research. *Biometrical Journal*, 59(6):1204–1220.

Hardin, J., Schmiediche, H., and Carroll, R. (2003). The simulation extrapolation method for fitting generalized linear models with additive measurement error. *The Stata Journal*, 3(4):373–385.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

He, W., Xiong, J., and Yi, G. (2012). SIMEX R package for accelerated failure time models with covariate measurement error. *Journal of Statistical Software*, 46(1):1–14.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Hsieh, F., Tseng, Y., and Wang, J. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, 62(4):1037–1043.

Huong, P., Nur, D., Pham, H., and Branford, A. (2018). A modified two-stage approach for joint modelling of longitudinal and time-to-event data. *Journal of Statistical Computation and Simulation*, 88(17):3379–3398.

Ibrahim, J., Chen, M., and Sinha, D. (2001). *Bayesian survival analysis*. Springer, 1st edition.

Ibrahim, J., Chu, H., and Chen, L. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796–2801.

Ibrahim, J. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *TEST*, 18(1):1–43.

Kalbfleisch, J. and Prentice, R. (2002). *The statistical analysis of failure time data*. John Wiley & Sons.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):325–335.

Little, R. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, R. and Rubin, D. (2019). *Statistical analysis with missing data*. John Wiley & Sons, 3rd edition.

Mauff, K., Steyerberg, E., Kardys, I., Boersma, E., and Rizopoulos, D. (2020). Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. *Statistics and Computing*, 30:999–1014.

Murawska, M., Rizopoulos, D., and Lesaffre, E. (2012). A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies. *Journal of Probability and Statistics*, 2012:1–18.

Murphy, S. and van der Vaart, A. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465.

Murtaugh, P., Dickson, E., van Dam, G., Malinchoc, M., Grambsch, P., Langworthy, A., and Gips, C. (1994). Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, 20(1):126–134.

Muthén, B., Asparouhov, T., Boye, M., Hackshaw, M., and Naegeli, A. (2009). Applications of continuous-time survival in latent variable models for the analysis of oncology randomized clinical trial data using Mplus. Technical report, Muthén & Muthén.

Neuhaus, A., Augustin, T., Heumann, C., and Daumer, M. (2009). A review on joint models in biometrical research. *Journal of Statistical Theory and Practice*, 3(4):855—-868.

Papageorgiou, G., Mauff, K., Tomer, A., and Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual Review of Statistics and Its Application*, 6(1):223–240.

Pinheiro, J. and Bates, D. (2000). Linear mixed-effects models: basic concepts and examples. In *Mixed-effects models in S and S-Plus*, pages 3–56. Springer.

Pinheiro, J. and Bates, D. (2021). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-152, `https://cran.r-project.org/package=nlme`.

Pinheiro, J. and Chao, E. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):58–81.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: with applications in R*. Chapman & Hall/CRC, 1st edition.

Rizopoulos, D. (2018). *JM: Joint modeling of longitudinal and survival data*. R package version 1.4-8, `https://cran.r-project.org/package=JM`.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):637–654.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, 95(1):63–74.

Schuurman, N., Grasman, R., and Hamaker, E. (2016). A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, 51(2–3):185–206.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In *AIDS epidemiology*, pages 231–255. Springer.

Stan Development Team (2020). *rstan: R interface to Stan*. R package version 2.21.2, `http://mc-stan.org`.

Stefanski, L. and Cook, J. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256.

Sweeting, M. and Thompson, S. (2011). 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.

Taylor, J., Park, Y., Ankerst, D., Proust-Lima, C., William, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213.

Therneau, T. (2011). *survival: Survival analysis*. R package version 3.2-11, `https://cran.r-project.org/package=survival`.

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Tsiatis, A., DeGruttola, V., and Wulfsohn, M. (1995). 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.

Verbeke, G. (1997). Linear mixed models for longitudinal data. In *Linear mixed models in practice*, pages 63–153. Springer.

Viviani, S., Alfó, M., and Rizopoulos, D. (2014). Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing*, 24(3):417–427.

Wang, P., Shen, W., and Boye, M. (2012). Joint modeling of longitudinal outcomes and survival using latent growth modeling approach in a mesothelioma trial. *Health Services and Outcomes Research Methodology*, 12(2–3):182–199.

Wu, L. (2009). *Mixed effects models for complex data*. Chapman & Hall/CRC, 1st edition.

Wu, L., Liu, W., and Hu, X. (2010). Joint inference on HIV viral dynamics and immune suppression in presence of measurement errors. *Biometrics*, 66(2):327–335.

Wu, L., Wei, L., Yi, G., and Huang, Y. (2011). Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, 2012:1–17.

Wu, L. and Yu, T. (2016). Joint modeling of longitudinal and survival data. In *Wiley StatsRef: Statistics Reference Online*, pages 1–9. American Cancer Society.

Wu, M. and Carroll, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):175–188.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

Yang, Y., Tong, T., and Li, G. (2019). SIMEX estimation for single-index model with covariate measurement error. *AStA Advances in Statistical Analysis*, 103(1):137–161.

Ye, W., Lin, X., and Taylor, J. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data – a two-stage regression calibration approach. *Biometrics*, 64(4):1238–1246.

Yu, M., Law, N., Taylor, J., and Sandler, H. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, 14(3):835–862.