Contributions on Joint Modeling of Sequential Times to Event with Longitudinal Information

Author: Natàlia Pallarès Fontanet
Advisor: Carles Serrat Piè
Department: Matemàtica Aplicada I
University: Universitat Politècnica de Catalunya
Contributions on Joint Modeling of Sequential Times to Event with Longitudinal Information

Author: Natàlia Pallarès Fontanet
Advisor: Carles Serrat Piè

Departament de Matemàtica Aplicada I,
Universitat Politècnica de Catalunya

October 2014
Abstract

In longitudinal studies it is often of interest to measure the association between a longitudinal marker repeatedly measured in time and the risk for an event. By postulating a model for the joint distribution, we acknowledge this link and we can assess the association between the longitudinal and the survival processes. In this work, we present the joint modeling approach and focus on the extension of the established methodology to include sequential survival times and more than one longitudinal variable.

Thus, on the one hand, the purpose of the present work is to introduce the methodology for the joint models (comparing the frequentist and Bayesian approaches) and to develop a new theory to include more than one survival time and more than one longitudinal variable. On the other hand, we study in depth the structure of the R packages that fit joint models (JM and JMbayes) for one survival time and for one longitudinal variable in order to give some hints on how they should be modified to implement the problem we have theoretically formulated.

First of all, we present the methodology for joint modeling for one survival time and one longitudinal variable from frequentist and Bayesian perspectives and illustrate how both approaches work with a real case example. Then, this work introduces a new methodology to deal with a joint model for $K > 1$ sequential survival times and $L > 1$ longitudinal variables, taking into account their particularities (such as the possible correlation that can arise between successive sequential times or between some of the longitudinal variables). After developing the methodology, we carry out an exhaustive study of the estimation functions in the two aforesaid R packages, point out their main differences and present which parts of the implementation should be modified and how to do it to incorporate $K$ sequential times and $L$ longitudinal variables.

The statistical software R has been used for all computations in this work. The packages that have been used (among others) are: nlme (Pinheiro et al., 2012), survival (Therneau, 2013), JM (Rizopoulos, 2013) and JMbayes (Rizopoulos, 2014a).

Keywords: joint models, sequential times, survival analysis, time-dependent covariates.
Resumen

En estudios longitudinales a menudo el interés radica en medir la asociación entre un marcador longitudinal repetidamente evaluado al largo del tiempo y el riesgo para un evento. Formulando un modelo para la distribución conjunta, podemos tener en cuenta esta relación y evaluar la asociación entre el proceso longitudinal y el proceso de supervivencia. En este trabajo, presentamos una aproximación del tipo joint modeling y nos centramos en extender la metodología existente para incluir tiempos de supervivencia secuenciales y más de una variable longitudinal.

Por lo tanto, por una parte el objetivo de este trabajo es introducir la metodología para los joint models (comparando las aproximaciones frequentista y bayesiana) y desarrollar una nueva teoría para incluir más de un tiempo de supervivencia y más de una variable longitudinal. Por otra parte, se profundiza en el estudio de la estructura de los paquetes de R que ajustan joint models (JM y JMbayes) para un tiempo de supervivencia y una variable longitudinal con el objetivo de dar indicaciones sobre cómo se tendrían que modificar para implementar el problema que se ha desarrollado teóricamente.

Primero de todo, presentamos la metodología para el caso de un tiempo de supervivencia y una variable longitudinal desde puntos de vista frequentista y bayesiano y ilustramos como funcionan con datos de un ejemplo real. Después, este trabajo introduce la nueva metodología para tratar con un joint model con $K > 1$ tiempos secuenciales y $L > 1$ variables longitudinales, teniendo en cuenta sus particularidades (como la posible correlación entre tiempos secuenciales sucesivos o entre diferentes variables longitudinales). Después de desarrollar la metodología, llevamos a cabo un estudio exhaustivo de las funciones de estimación de los paquetes mencionados anteriormente, señalamos las principales diferencias y detectamos qué partes de la implementación se tendrían que modificar y cómo hacerlo para incorporar $K$ tiempos de supervivencia secuenciales y $L$ variables longitudinales.

El software estadístico R ha sido utilizado para todos los cálculos de este trabajo. Entre otros, los paquetes usados han sido: nlme (Pinheiro et al., 2012), survival (Therneau, 2013), JM (Rizopoulos, 2013) y JMbayes (Rizopoulos, 2014a).

Palabras clave: joint models, tiempos secuenciales, análisis de la supervivencia, time-dependent covariates.
Resum

En estudis longitudinals, sovint interessa mesurar l’associació entre un marcador longitudinal mesurat al llarg del temps i el risc d’experimentar un cert esdeveniment. Formulant un model per a la distribució conjunta de les dades longitudinals i les dades de supervivència, podem tenir en compte aquest relació i avaluar l’associació entre els dos processos. En aquest treball, presentem l’aproximació de tipus joint modeling i ens centrem a estendre-la per incloure-hi temps de supervivència seqüencials i més d’una variable longitudinal.

Per tant, per una banda, l’objectiu d’aquest treball és introduir la metodologia per als joint models (comparant les aproximacions freqüentista i bayesiana) i desenvolupar una nova teoria per incloure més d’un temps de supervivència i més d’una variable longitudinal. D’altra banda, s’aprofundeix en l’estudi de l’estructura dels paquets de R que ajusten joint models (JM i JMbayes) per a un temps de supervivència i una variable longitudinal amb l’objectiu de donar indicacions sobre com s’haurien de modificar per implementar el problema que s’ha formulat teòricament.

En primer lloc, presentem la metodologia pel joint modeling d’un temps de supervivència i una variable longitudinal des dels punts de vista freqüentista i bayesià i il·lustrem com funcionen les dues aproximacions amb dades d’un exemple real. A continuació, introduïm una nova metodologia per tractar el cas de $K > 1$ temps de supervivència seqüencials i $L > 1$ variables longitudinals, tenint en compte les particularitats d’aquest tipus de dades (com lapossible correlació que pot existir entre temps seqüencials successius o entre algunes de les variables longitudinals). Després de desenvolupar la metodologia, portem a terme un estudi exhaustiu de les funcions d’estimació dels paquets esmentats anteriorment, assenyalem les principals diferències i presentem quines parts de la implementació s’haurien de modificar i com fer-ho per incorporar $K$ temps seqüencials i $L$ variables longitudinals.

S’ha utilitzat el software estadístic R per a tots els càlculs en aquest treball. Entre d’altres, els paquets que s’han fet servir són: nlme (Pinheiro et al., 2012), survival (Therneau, 2013), JM (Rizopoulos, 2013) i JMbayes (Rizopoulos, 2014a).

Paraules clau: joint models, temps seqüencials, anàlisi de la supervivència, time-dependent covariates.
# Contents

Abstract i

Resumen iii

Resum v

List of Figures ix

List of Tables x

1 Introduction and goals 1

2 Longitudinal and survival models 3
   2.1 Introduction .......................................................... 3
   2.2 Longitudinal data analysis ......................................... 3
       2.2.1 Variability in longitudinal data ............................... 4
       2.2.2 The linear mixed model ....................................... 5
   2.3 Survival Data Analysis ............................................ 7
       2.3.1 Basic functions on survival analysis .......................... 8
       2.3.2 The proportional hazards Cox Model .......................... 8
       2.3.3 Extended Cox Model for time-dependent covariates ........ 9
   2.4 Motivating dataset: The TIBET clinical trial .................... 10

3 Joint modeling framework 13
   3.1 Introduction .......................................................... 13
   3.2 Longitudinal submodel ............................................. 13
   3.3 Survival submodel ................................................ 14
   3.4 Joint Model formulation .......................................... 15
   3.5 Joint Model estimation ............................................ 15
   3.6 Joint Model illustration: TIBET ................................. 17
       3.6.1 Models for time to treatment reinitiation $T_1$ ............ 18
       3.6.2 Models for time to treatment interruption $T_1 + T_2$ .... 20

4 Methodology JMseq($K, L$) 23
   4.1 Introduction .......................................................... 23
   4.2 $K$ survival times and one longitudinal variable ............... 24
       4.2.1 $K = 2, L = 1$: JMseq(2, 1) ............................... 24
       4.2.2 $L = 1$: JMseq($K, 1$) ..................................... 28
   4.3 One survival time and $L$ longitudinal variables ............... 30
       4.3.1 $K = 1, L = 2$: JMseq(1, 2) ............................... 31
| 4.3.2 | \( K = 1 \): JMseq(1, L) | 32 |
| 4.4 | \( K \) sequential times and \( L \) longitudinal variables | 33 |
| 4.4.1 | \( K = 2, L = 2 \): JMseq(2, 2) | 33 |
| 4.4.2 | JMseq(\( K, L \)) | 35 |
| 5 | Towards JMseq(\( K, L \)) algorithm: packages JM and JMbayes | 39 |
| 5.1 | Introduction | 39 |
| 5.2 | Package JM | 40 |
| 5.2.1 | Base structure | 40 |
| 5.2.2 | Most important procedures | 41 |
| 5.3 | Package JMbayes | 43 |
| 5.3.1 | Base structure | 43 |
| 5.3.2 | Most important procedures | 45 |
| 5.4 | Similarities and differences between JM and JMbayes | 46 |
| 5.4.1 | Baseline risk function | 46 |
| 5.4.2 | Linear predictor for the survival submodel | 47 |
| 5.4.3 | Structure differences | 48 |
| 5.5 | Extension towards JMseq(\( K, L \)) | 49 |
| 6 | Discussion and future research | 53 |
| 6.1 | Discussion and conclusions | 53 |
| 6.2 | Future research | 54 |
| Bibliography | | 55 |
| A | R code | 58 |
# List of Figures

2.1 Sources of variation in longitudinal data .......................... 5  
2.2 Profiles of treatment during the TIBET study ..................... 11  
2.3 Evolution of $\sqrt{CD4}$ and log (viral load) for a particular subject ......................................... 12  
4.1 Types of observed data ................................................. 26  
4.2 Construction of the longitudinal variable with sequential times ........................................ 30
List of Tables

3.1 Joint models for time to reinitiation ($T_1$) adjusting for the longitudinal variable $\sqrt{CD4}$ ................................................................. 19
3.2 Joint models for time to reinitiation ($T_1$) adjusting for the longitudinal variable log (VL) ................................................................. 20
3.3 Joint models for time to withdrawal ($T_1 + T_2$) adjusting for the longitudinal variable $\sqrt{CD4}$ ................................................................. 21
3.4 Joint models for time to withdrawal ($T_1 + T_2$) adjusting for the longitudinal variable log (VL) ................................................................. 22
4.1 Dimension and number of parameters for the covariance matrix of the random effects ................................................................. 38
Chapter 1

Introduction and goals

Joint modeling of longitudinal and time-to-event data is an active area of biostatistics that has received a lot of attention in the recent years mainly because in longitudinal studies the interest often lies in the relation between a longitudinally measured marker and a time-to-event outcome. Some examples of this relation are studies on human immunodeficiency virus (HIV) where we are interested in the association between the longitudinal CD4 cell count measurements and the time to death (Tsiatis et al., 1995) or prostate cancer studies in which the practitioner is interested in associating the longitudinal level of prostate specific antigen (PSA) measurements with the time to cancer recurrence (Proust-Lima and Taylor, 2009).

One of the main reasons for the increasing interest in this area is that joint models are applicable in very different situations. For example:

1. When the interest is on the survival outcome and we wish to account for the effect of a longitudinal variable as a time-dependent covariate. Approaches such as the Cox proportional hazards model are not applicable when the covariate is endogenous and time-dependent (Kleinbaum and Klein, 2005). By postulating a model for the joint distribution, we acknowledge this link and we obtain a more accurate estimate for the association between the two processes.

2. When the focus is on the longitudinal outcome and we wish to correct for the potential dropout due to not having longitudinal measurements available at and after the event time. In order to obtain valid inferences, we need to postulate a model for the joint distribution of the longitudinal and missingness processes.

In both settings, the joint distribution of the event times and the longitudinal measurements is modeled through a set of random effects that are assumed to account for the association between the two outcomes. Both Tsiatis and Davidian (2004) and Yu et al. (2004) give excellent overviews of this area.

Up to now, most of the work in joint models has focused on models with a single longitudinal outcome associated with the survival times. However, in many longitudinal studies, patients are repeatedly measured, generating a series of outcomes that are potentially predictive for the time until the event. As these longitudinal outcomes are examples of endogenous time-dependent covariates measured with error, a joint modeling approach is required. In this regard, there is not much literature on methodologies able to include in the joint models not one but more than one longitudinal outcomes. Moreover,
it can be also of importance to study the case where each individual has more than one
time to event, that is, when there are two or more events of interest. Although this
situation comprehends different types of data such as sequential, parallel or family data,
in this work we will focus in sequential times. Huertas-Campos (2011) developed the
methodology for the case of two survival times and two longitudinal variables but more
complex cases have not been studied yet.

What we aim in this work is to study the case of not only one survival time and one
longitudinal variable but also the case in which there are more than one survival time
(for example sequential times for each individual) and more than one longitudinal variable
(in order to have more information to help model the time until the event). Moreover,
as the main importance of the theory we develop in this work is its application to real
cases, we take an in depth look to the two main R packages to estimate joint models in
the case of one longitudinal outcome and one event of interest: JM (Rizopoulos, 2010)
and JMbayes (Rizopoulos, 2014a) in order to detect which parts of the code need to be
modified and how to do so to take into account the case for several longitudinal variables
and several time-to-event times.

The purpose of the present work is, then, threefold: on the one hand, after introducing the
methodology for the joint models, we develop a new theory for the case of a joint model
with more than one survival time and more than one longitudinal variable. On the other
hand, we compare the frequentist and Bayesian approaches for estimating the parameters
in these models. Last, we study in depth the structure of the two aforesaid R packages
that are used to estimate joint models with one survival time and one longitudinal variable
in order to understand how they work and give some hints on how they should be modified
to implement an extension of the problem we have theoretically formulated.

The rest of this master’s degree thesis is structured as follows: in Chapter 2, the building
blocks of joint models (relative risk models for survival data, following Cox (1972)
and Kleinbaum and Klein (2005), and linear mixed-effects model for longitudinal data,
following Verbeke and Molenberghs (2000)) are introduced. This chapter also contains a
description of the dataset that motivated the development of a methodology able to work
with several survival times and several longitudinal variables. Chapter 3 introduces the
standard joint model and presents both maximum likelihood and Bayesian estimation.
It also contains an illustration of how packages JM and JMbayes work. In Chapter 4
the methodology for the case of $K$ sequential times and $L$ longitudinal variables, which
will be denoted as JMseq($K, L$), is developed. We first study the case of more than one
sequential time and just one longitudinal variable. Then, we focus on one survival time
and introduce more than one longitudinal variable. At the end of this Chapter 4, we arrive
at our main goal which is the study of the case of $K$ sequential times and $L$ longitudinal
variables. The aim of Chapter 5 is to carry out a comprehensive analysis of packages JM
and JMbayes, studying their structure, the similarities and differences between them and
giving indications on how the mentioned packages could be extended for the case of more
than one survival time and more than one longitudinal variable.

The statistical software R has been used for all the computations in this work.
Chapter 2

Longitudinal and survival models

2.1 Introduction

This chapter introduces both the model for the longitudinal data and the model for the survival data. For the longitudinal information, we construct a linear mixed-effects model for the analysis of continuous longitudinal responses which is the first building block of joint models for longitudinal and time-to-event data. We describe the variability that can be found in a set of longitudinal data and how the parameters of the model are estimated.

Secondly, we introduce the basic concepts for the analysis of survival data as well as relative risk models. We go through some models for time-to-event data arriving to the extended Cox Model, the type of relative risk model that constitutes the second building block for the joint models we are constructing. We focus on the handling of time-dependent covariates, presenting both endogenous and exogenous covariates.

Because the separate analysis of longitudinal and survival data may lead to inefficient or biased results, in next Chapter we will model them jointly in order to incorporate all the information simultaneously.

In order to study the relation between survival times and longitudinal information, at the end of the Chapter we present the TIBET clinical trial, which is the motivating dataset for this work. The TIBET clinical trial contemplated the incorporation of interruption periods in the administration of an intensive therapy for HIV infected patients. The time that a patient needs before restarting or suspending treatment is of clinical interest. Because we have the information of previous lifetimes and the evolution of longitudinal markers, it is plausible to think of a joint model for sequential times to event and longitudinal information.

2.2 Longitudinal data analysis

Following Verbeke and Molenberghs (2000), correlated data includes several multivariate structures such as clustered data, spatial correlated data, repeated measurements or longitudinal data. We will focus on longitudinal data, which can be defined as data resulting from observations of subjects that are measured repeatedly over time. The main characteristics of longitudinal data are: 1) the outcome is measured repeatedly
within a set of units, 2) longitudinal measures are positively correlated within subjects, 3) repeated measures from a single subject allow us to capture within-subject patterns of change and 4) the number of observations and time points can vary from one subject to another leading to unbalanced longitudinal designs.

With this description, it is clear that in longitudinal setting we expect the repeated measurements taken on the same subject to be correlated. Thus, standard statistical tools, such as t-test or simple linear regression, that assume independent observations, are not appropriate for the analysis of this kind of data.

2.2.1 Variability in longitudinal data

When working with longitudinal data it can not be assumed that measurements are independent, because as the individuals are followed along time their measurements are highly likely to be correlated. Thus, when modeling them, we have to take into account that part of the variability of the data is produced by each individual by itself. Moreover, because we compare the measurements from different subjects on the population, we have to take into account the variability between them. Thus, there are two different sources of variability in longitudinal data:

**Between-subject variability**

It represents the variation of the subject-specific mean profile with respect to the population mean profile. This variability is based on the underlying subject-specific behaviour (due to, for example, genetic or social factors) and how its trend can be derived from all the repeated measurements that have been recorded.

**Within-subject variability**

It measures the variability in response within the same subject. This variation is based on the variability around the subject’s true and unobservable marker, because each marker can be subject to a potential random measurement error.

In order to illustrate these variabilities, in Figure 2.1 we have the representation of sources of variability in a balanced study with three individuals and six measurements for each individual. The black points represent the measurements of the variable of interest at each time point. They are subject to measurement error since direct observation without adverse effects is difficult or impossible. The dotted lines stand for the unobservable true values of the response. The black solid lines represent the subject-specific evolutions, whereas the red solid line is the population-average evolution.
2.2.2 The linear mixed model

The idea behind longitudinal data is that each individual of the population has his own subject-specific mean profile over time. This, coupled with the averaged evolution, describes the trend of each subject. Thus, in longitudinal data analysis we have:

- **Fixed-effects component** that captures the average evolution in time of a variable. This average is an estimate of the evolution of the covariate in the target population.

- **Random-effects component** that captures the particular evolution in time for each of the individuals under study. This part takes into account that the data from an individual are correlated.

To describe separately the data of each subject, a simple linear regression model with an intercept and a linear time effect may seem adequate. However, as different subjects might have different values for intercepts and slopes we must reformulate the model including random effects in the intercept, the slope or both the intercept and the slope to account for this variability coming from each subject.

Formally, suppose that we have \( n \) subjects, each of them with a different number \( n_i \) of measurements of the variable of interest, \( i = 1, \ldots, n \). These measurements are taken at
different time points. Let $y_{ij}$ be the response variable of the $i$-th individual observed at $t_{ij}$, with $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. This $t_{ij}$ are called the calendar times. So, for the $i$-th subject, its outcomes can be expressed as a $n_i$-dimensional vector $y_i = (y_{i1}, \ldots, y_{in_i})^T$. Assuming the longitudinal outcome is normally distributed, the general linear mixed model can be written as (Laird and Ware, 1982; Verbeke and Molenberghs, 2000):

\[
\begin{align*}
  y_i &= X_i \beta + Z_i b_i + \epsilon_i, \\
  b_i &\sim N(0, D), \\
  \epsilon_i &\sim N(0, \sigma^2),
\end{align*}
\]  

(2.1)

where $X_i$ is the design matrix for the fixed-effects regression coefficients $\beta$ and $Z_i$ is the design matrix for the random-effects regression coefficients $b_i$. Thus, $\beta$ denotes the vector of the unknown fixed effects and $b_i$ the vector for the random effects in the model with $D$ being its covariance matrix. Moreover, $\epsilon_i$ is the random error and $\sigma^2$ represents the within-subject variation. Thus, we assume random effects are normally distributed with mean zero and variance-covariance matrix $D$. Moreover, we assume that they are independent of the error terms $\epsilon_i$, that is, \{b_1, \ldots, b_n\} independent of \{\epsilon_1, \ldots, \epsilon_n\} and that errors of an individual are independent between them given $b_i$.

The interpretation of $\beta$ is the same as in a linear regression model, and $b_i$ can be interpreted as how the regression parameters for the $i$-th subject deviate from the population’s parameters.

**Estimation of the parameters**

The estimation of the parameters of the linear mixed model is based on maximum likelihood. In particular, for the $i$-th individual, the marginal density of the observed data is given by

\[
p(y_i) = \int p(y_i | b_i) p(b_i) \, db_i.
\]  

(2.2)

Because both the conditional distribution of the longitudinal responses given the random effects and the distribution of the random effects are normal, the integral in (2.2) has a closed-form solution $y_i \sim N(X_i \beta, V_i)$, where $V_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$. Let $\alpha$ denote the vector of all variance and covariance parameters found in $V_i$, that is the elements in $D$ and the parameters in $\sigma^2$.

If we assume that $\alpha$ is known, the maximum likelihood estimator of the fixed-effects vector $\beta$ can be obtained by the generalized least squares estimator (Laird and Ware, 1982)

\[
\hat{\beta} = \left( \sum_{i=1}^n X_i^T \hat{W}_i X_i \right)^{-1} \sum_{i=1}^n X_i^T \hat{W}_i y_i,
\]  

(2.3)

where $W_i = V_i^{-1}$. 
Estimates for the standard errors of the fixed-effects can be obtained calculating the variance of the estimator

$$\hat{\text{Var}}(\hat{\beta}) = \left( \sum_{i=1}^{n} X_i^T \hat{V}_i^{-1} X_i \right)^{-1}. \quad (2.4)$$

Predictions for the subject-specific random effects $b_i$ are obtained by the best linear unbiased predictor (BLUP)

$$\hat{b}_i = \hat{D}Z_i^T \hat{V}_i^{-1} \left( y_i - X_i \hat{\beta} \right), \quad (2.5)$$

where $\hat{D}$ and $\hat{V}_i$ are the restricted maximum likelihood (REML) estimators.

When $\alpha$ is not known, but an estimate $\hat{\alpha}$ is available, we can set $V_i = \hat{V}_i(\hat{\alpha}) = \hat{W}_i^{-1}$ and estimate $\beta$ by replacing $W_i$ by $\hat{W}_i$ in (2.3). To estimate $\alpha$, the two frequently used methods are maximum likelihood estimation and restricted maximum likelihood estimation.

### 2.3 Survival Data Analysis

The aim of survival data analysis is to analyse the time until an event of interest occurs. The response variable, $T$, is the time until that event, often called survival time or failure time. The first feature that must be taken into account when working with survival data is the shape of its distribution. Since event times must be positive, we usually have skewed shape distributions that make the assumptions on normality do not hold.

Another characteristic that distinguishes survival times is censoring. That means that the event of interest is not observed on all subjects under study. There are different types of censoring such as left censoring, right censoring or interval censoring. In this work we are going to consider only right censoring, that occurs when the subject has not yet experienced the event of interest at the time when the follow-up period ends. In these cases, the only information that we have about the true survival time is that it is larger than the observed survival time.

Denoting $T$ a positive random variable representing the observed survival time, under the presence of right censoring, each individual provides a true survival time $T_i^*$ and a right-censored time $C_i$. With these measures, each individual can be summarized by the observed survival time $T_i = \min\{T_i^*, C_i\}$ and by the event indicator $\delta_i = I(T_i^* \leq C_i)$, thus, for each individual we have its information collected as $(T_i, \delta_i)$.

Considering the skewed shapes and the censoring mechanism, standard statistical tools cannot be used because they assume that we have complete information, which is not the case in survival data analysis.
2.3.1 Basic functions on survival analysis

We assume \( T^* \) is a continuous random variable that represents the time until some event. Let \( F(t) \) be its cumulative distribution function and \( f(t) \) its probability density function. The interesting functions in survival analysis are:

- The survival function, \( S(t) \), which is the probability that the event of interest has not yet occurred at time \( t \) and it is written as
  \[
  S(t) = P(T^* > t). \tag{2.6}
  \]

- The hazard function, \( h(t) \), that denotes the rate of occurrence of the event at a given time \( t \), given that subjects are still at risk, can be written as
  \[
  h(t) = \lim_{\Delta t \to 0^+} \frac{P(t < T^* < t + \Delta t | T^* > t)}{\Delta t}. \tag{2.7}
  \]

It can be proved that

\[
  h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log(S(t)). \tag{2.8}
  \]

2.3.2 The proportional hazards Cox Model

The proportional hazards Cox Model (Cox, 1972) gives an expression for the hazard for an individual at time \( t \) with a given set of time-independent explanatory covariates \( \mathbf{w}_i \) (called baseline covariates) and it can be written as

\[
  h_i(t|\mathbf{w}_i) = h_0(t)\psi(\mathbf{w}_i), \quad t \geq 0, \tag{2.9}
  \]

where \( h_0(t) \) is the baseline hazard function that represents the hazard function when \( \mathbf{w}_i = 0 \) and is typically left unspecified. Given that \( \mathbf{w}_i = (w_{i1}, \ldots, w_{ip})^T \) is a \( p \)-dimensional vector of time-independent covariates for the \( i \)-th subject, \( \psi(\mathbf{w}_i) \) is a non-negative function that contains the information about the effects of the explanatory covariates for the \( i \)-th subject. Usually, we take \( \psi(\mathbf{w}_i) = \exp(\gamma^T\mathbf{w}_i) \), where \( \gamma = (\gamma_1, \ldots, \gamma_p)^T \) is an unknown \( p \)-dimensional parameter vector. So, the expression (2.9) becomes

\[
  h_i(t|\mathbf{w}_i) = h_0(t) \exp\{\gamma_1 w_{i1} + \cdots + \gamma_p w_{ip}\} = h_0(t) \exp\left\{\sum_{k=1}^{p} \gamma_k w_{ik}\right\}. \tag{2.10}
  \]

This model is called proportional hazards model because if we consider two subjects \( i \) and \( j \) with covariates \( \mathbf{w}_i \) and \( \mathbf{w}_j \), their hazard ratio is constant and given by

\[
  \frac{h_i(t|\mathbf{w}_i)}{h_i(t|\mathbf{w}_j)} = \frac{h_0(t) \exp\{\sum_{k=1}^{p} \gamma_k w_{ik}\}}{h_0(t) \exp\{\sum_{k=1}^{p} \gamma_k w_{jk}\}} = \exp\left\{\sum_{k=1}^{p} \gamma_k (w_{ik} - w_{jk})\right\}, \tag{2.11}
  \]

and, therefore, their hazard rates are proportional and are not time-dependent.
2.3.3 Extended Cox Model for time-dependent covariates

The relative risk model that we have introduced assumes that the hazard depends only on baseline covariates whose values are constant during follow-up. However, in many studies the interest can be in studying whether time-dependent covariates are associated with the risk of experiencing the event. Before explaining how this covariates need to be handled, we are going to distinguish between two different types of time-dependent covariates, endogenous (or internal) and exogenous (or external). The need for this distinction is that the endogenous variables require special treatment compared to the exogenous ones.

Types of time-dependent covariates

Denoting \( y_i(t) \) the time-dependent covariates at time \( t \) for the \( i \)-th individual, and \( \mathcal{Y}_i(t) = \{ y_i(s), 0 < s \leq t \} \) the covariate history for this individual up to time \( t \), there are two types of time-dependent covariates

1. **External** or exogenous covariates: following Kalbfleisch and Prentice (2002), these are covariates that satisfy the relation

\[
P(s \leq T_i^* < s + ds | T_i^* \geq s, \mathcal{Y}_i(t)) = P(s \leq T_i^* < s + ds | T_i^* \geq s, \mathcal{Y}_i(t)) \tag{2.12}
\]

for all \( s, t \) such that \( 0 < s \leq t \) and \( ds \to 0 \). An exogenous covariate is a predictable process, so its value at any time \( t \) is known infinitesimally before \( t \). In other words, its future path up to time \( t > s \) is not affected by the occurrence of failure at time \( s \). An example of an exogenous covariate is the time of the day or the season of the year. Another type of external covariates are processes external to the subject under study. For instance, the level of air pollution, that may be associated with the frequency of asthma attacks. For external covariates, we can use the relation between the survival function and the hazard function to directly define the survival function conditional on the covariate as follows

\[
S_i(t | \mathcal{Y}_i(t)) = P(T^* > t | \mathcal{Y}_i(t)) = \exp \left\{ - \int_0^t h_i(s | \mathcal{Y}_i(s)) \, ds \right\}. \tag{2.13}
\]

2. **Internal** or endogenous covariates: this are covariates that arise when there are time-dependent measurements taken on the subjects under study. The change of the covariate depends on the individual. This type of covariates are usually measured with error and it is reasonable to assume that the observed marker levels are a contaminated version of the true marker levels. Examples of endogenous covariates include biomarkers such as CD4 cell counts for HIV-infected patients or serum bilirubin levels for patients with primary biliary cirrhosis.

Extended Cox Model

Proportional hazards (PH) model assumed that the value of covariates is constant over time. However, relevant covariates for survival analysis may change in the observation period. Following Kleinbaum and Klein (2005), we can write the extended Cox Model to handle both time-independent and time-dependent covariates.

If \( y_i(t) \) is the covariate at time \( t \) for the \( i \)-th subject, \( \mathcal{Y}_i(t) = \{ y_i(s), 0 \leq s \leq t \} \) is the covariate history until \( t \) and \( \mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{ip})^T \) is the vector of baseline covariates for the \( i \)-th subject then the extended Cox Model can be written as
\[ h_i(t|\mathbf{y}_i(t), \mathbf{w}_i) = h_0(t) \exp\{\gamma^T \mathbf{w}_i + \alpha y_i(t)\}. \] (2.14)

In a similar way as in the PH Cox Model, the extended model contains a baseline hazard function multiplied by an exponential part. However, in the extended Cox Model, the exponential part contains both time-independent and time-dependent predictors. In this model the regression coefficients \( \gamma \) and \( \alpha \) have the same interpretation as the coefficients in the proportional hazards Cox Model.

Despite having more flexibility than the PH Cox Model, the extended Cox Model is not appropriate when the time-dependent covariates are endogenous variables, because for these variables we do not have the complete knowledge of the covariate history for all individuals while on study. In the extended Cox Model time-dependent covariates are assumed to change its value at follow-up visits and to remain constant in the interval between them. This assumption is unrealistic when having endogenous covariates, such as biomarkers, because it is not reasonable to think they remain constant between measurement points. This assumption lead to biased estimations. As a consequence, it is necessary to introduce a modeling framework to deal with the special features of endogenous time-dependent covariates.

### 2.4 Motivating dataset: The TIBET clinical trial

TIBET was a clinical trial carried out by the Fundació Lluita contra la Sida at Hospital Germans Trias i Pujol of Badalona (Spain) that contemplated the incorporation of interruption periods in the administration of HAART (Highly Active AntiRetroviral Therapies) for HIV infected patients (Ruiz et al., 2007). The aim of the study was to analyze the time that a patient needs before restarting or suspending treatment given the values of biological markers recorded in the follow-up period. This clinical trial motivated the proposal of a joint model to analyze sequential times to event with more than one longitudinal variable.

Patients who entered the study were randomized to continue the HAART therapy \((n = 101)\) or to follow a protocol of treatment-interruption \((n = 100)\). Patients who went under the treatment-interruption protocol started with suspension of treatment (stage OFF) and every four weeks information on CD4 cell counts (which are cells that send signals to activate the body’s immune response) and viral load (which is the level of HIV in a patient’s blood) was registered. If the conditions of the patient had deteriorated, the therapy was restarted (stage ON). In Figure 2.2 we can see different cycles of treatment recorded in the study. Patients in the control group received standard HAART during the follow-up, whereas subjects interrupting therapy could experiment different number of OFF-ON and ON-OFF changes. Numbers at the right of the figure represent the number of patients according to the interruption-reinitation sequence.

Therefore, the treatment generates a sequence of stages OFF-ON-OFF-ON that defines a sequence of times \(T_1, T_2, \ldots, T_K\) that are the length that a patient stays on each stage. That means that \(T_1\) is the duration of the first stage OFF (time, from randomization, that a patient stays without therapy), \(T_2\) is the duration of the first stage ON (or, equivalently, the time since the patient starts treatment until he is switched off) and \(T_3\) will be the second time OFF (the time the patient is without treatment for the second time). This
Chapter 2. Longitudinal and survival models

\( \{T_i : i = 1, \ldots, K\} \) occur sequentially and in order, meaning that transition \( l \) (moving from stage OFF to stage ON or vice versa) only occurs after the previous \( l - 1 \) transitions have occurred. One of the goals of this type of clinical trial was to characterize the time spent in the first state ON \( (T_2) \) taking into account the information of the first state OFF \( (T_1) \).

Together with these times, there were records of the CD4 cell count and of the viral load for each patient along time. These two biological markers can be the longitudinal variables included in the joint model for modeling the survival times with more than one longitudinal variable. First of all we need to model them along \( T_1 \) and \( T_2 \) (or \( T_1, \ldots, T_K \) if we extend the case to \( K \) sequential times). This longitudinal information can follow different trends such as linear, parabolic or piecewise.

**Figure 2.2:** Profiles of treatment interruption and reinitiation during the TIBET study. The patients followed different cycles of treatment withdrawal (black) and reinitiation (white). Source: Ruiz et al. (2007).

The conditions for reinitiating the therapy (moving from stage OFF to stage ON) were either

- The CD4 levels decrease to less than 350 cell/mm\(^3\).
- The viral load increases to more than 100000 copies/ml.
- Development of an AIDS-defining event.

Analogously, a patient moved from stage ON to stage OFF (suspending the therapy) when the two following conditions held

- The CD4 levels reach more than 500 cell/mm\(^3\).
- The viral load decreases to undetectable levels (such as less than 50 copies/ml).

In Figure 2.3 we can see the evolution of the CD4 levels and the viral load for one particular subject who followed the protocol of treatment-interruption. Note that we have represented the square root of the CD4 cell counts and the logarithm of the viral load because these are the transformed variables that we will include in our models. Jointly with the evolution of the longitudinal variables we can see the changes in stages ON and OFF. We observe that in ON periods, the CD4 cell counts go up and the viral
load goes down. The opposite happens in OFF stages. Moreover, we have represented the limit values for reinitiating and suspending treatment for both CD4 and viral load.

![Graph showing longitudinal response of CD4 and log(viral load) with stages ON-OFF.](image)

**Figure 2.3:** Evolution of $\sqrt{\text{CD4}}$ and log (viral load) for a particular subject jointly with the sequence of stages ON-OFF. The solid lines stand for the conditions for reinitiating treatment (green ones) and interrupting treatment (red).

In this case, and because a latent association between survival and longitudinal variables existed, a joint model must be fitted in order to estimate properly both processes and thus, since we have sequential times and two longitudinal variables, a joint model for sequential times to event and more than one longitudinal variable is needed.
Chapter 3

Joint modeling framework

3.1 Introduction

The motivating idea behind the joint models is to couple survival data, which is of primary interest, and a suitable model for the repeated measurements of the endogenous covariate. Joint models were introduced during the 90’s (Tsiatis et al., 1995; Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997) and since then they have been applied to a lot of studies in epidemiological and biomedical areas. Recently, Rizopoulos has made a great contribution to the joint modeling methodology, first by an overview of the theory and applications of joint modeling (Rizopoulos, 2012) and secondly by developing the JM (Rizopoulos, 2010) and JMbayes (Rizopoulos, 2014a) R packages for the frequentist and Bayesian approaches, respectively. In this line, joint modeling also illustrates how frequentist and Bayesian statistics can be combined to reach a complex goal, using the strengths of both approaches.

In this Chapter, based on the linear mixed-effects model and the relative risk model introduced in Chapter 2, we present the standard joint model for longitudinal and survival data. A typical joint model setting is to assume a linear mixed-effects model for the longitudinal covariates and a Cox model or an accelerated failure time model for the survival data, with the two models sharing some random effects or covariates.

We discuss also maximum likelihood estimation of the parameters from the frequentist and Bayesian points of view.

3.2 Longitudinal submodel

Because the longitudinal covariate is an endogenous time-dependent covariate that we have measured with error, we assume that the observed value of the longitudinal covariate \( y_i(t) \) equals the true and unobserved level of the endogenous variable at time \( t \), \( m_i(t) \), plus a random error term, \( \epsilon_i(t) \). If \( \mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\} \) is the history of the unobserved longitudinal process until \( t \), we need to estimate \( m_i(t) \) and reconstruct the complete longitudinal history \( \mathcal{M}_i(t) \) for each subject.

To describe the subject-specific time evolutions, we focus on normal data and use a linear mixed-effects model as
\[
\begin{align*}
\begin{cases}
  y_i(t) = m_i(t) + \epsilon_i(t), \\
  m_i(t) = x_i^T(t)\beta + z_i^T(t)b_i, \\
  b_i \sim \mathcal{N}(0, D), \\
  \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)
\end{cases}
\end{align*}
\] (3.1)

We should note that in (3.1), both the design vector \(x_i(t)\) for the fixed effects \(\beta\) and the design vector \(z_i(t)\) for the random effects \(b_i\) as well as the error term \(\epsilon_i(t)\) are time-dependent. This time structure and the use of subject-specific random effects allows us to reconstruct the complete path of the time-dependent process \(M_i(t)\).

Moreover, as in the previous Chapter, we assume that error terms are mutually independent, independent of the random effects and normally distributed with mean zero and variance \(\sigma^2\).

### 3.3 Survival submodel

To quantify the relation between \(m_i(t)\) and the risk for an event at time \(t\), we postulate a relative risk model for \(t > 0\)

\[
h_i(t|M_i(t), w_i) = h_0(t) \exp\{\gamma^Tw_i + \alpha m_i(t)\},
\] (3.2)

where, as we have said before, \(m_i(t)\) is the true and unobserved value of the longitudinal (endogenous) variable at time \(t\) and \(M_i(t)\) is the history of the unobserved longitudinal process until \(t\). To quantify the association between the true marker levels and the risk for an event we use \(\alpha\). The baseline risk function is represented by \(h_0(t)\), \(w_i\) is the vector of baseline covariates for the \(i\)-th subject and \(\gamma\) its corresponding vector of regression coefficients.

In survival analysis, the baseline risk function \(h_0\) is typically left unspecified, but in the joint modeling framework leaving this function unspecified leads to an underestimation of the standard error of the parameter estimates (Hsieh et al., 2006). Hence, to avoid these problems, to complete the definition of the model is necessary to discuss the choice for \(h_0\).

We could use a hazard function corresponding to a known parametric distribution such as the ones typically used in the context of survival analysis, Weibull, Gamma or log-normal. Alternatively, we can use a parametric but flexible specification such as step-functions and linear splines (Whittemore and Killer, 1986), B-splines approximation (Rosenberg, 1995) or restricted cubic splines (Herndon and Harrel, 1996). We formalize here two of the options, that are simple and work satisfactorily in practice: the piecewise-constant model and the regression splines approach.

For the piecewise-constant model, the baseline risk function takes the form of

\[
h_0(t) = \sum_{q=1}^{Q} \xi_q I(\nu_{q-1} < t < \nu_q),
\] (3.3)

where \(0 = \nu_0 < \nu_1 < \cdots < \nu_Q\) is a split of the time scale (with \(\nu_Q\) being larger than the largest observed time) and \(\xi_q\) denotes the value of the hazard in the interval \((\nu_{q-1}, \nu_q)\).
For the regression splines model, the logarithm of the baseline risk function is expanded into B-splines basis functions for cubic splines and can be expressed as

\[
\log h_0(t) = \gamma_{h,0} + \sum_{q=1}^{Q} \gamma_{h,q}B_q(t,\nu),
\]

where \(B_q(t,\nu)\) denotes the \(q\)-th basis function of a B-spline with knots \(\nu = (\nu_1, \ldots, \nu_Q)\) and \(\gamma_{h,0}\) is the vector of spline coefficients. Note that when the number of knots increases, the specification of the baseline hazard becomes more flexible.

In both approaches, after the number of knots have been chosen, their location is typically based on percentiles of either the observed event times \((T_i = \min(T_i^*, C_i))\) or only the true event times \(\{T_i : T_i^* \leq C_i, i = 1, \ldots, n\}\) to allow for more flexibility in the region with greatest density.

In fact, as we will see in Chapter 5, these two options that we have formalized are the options by default when using packages \texttt{JM} and \texttt{JMbayes} (piecewise for \texttt{JM} and regression splines for \texttt{JMbayes}).

### 3.4 Joint Model formulation

Taking into account the two submodels we have defined and explained in previous sections and their particularities, following Rizopoulos (2012) we are going to build the joint model by joining the two submodels, as follows

\[
\begin{cases}
    y_i(t) = m_i(t) + \epsilon_i(t) = \mathbf{x}_i^T(t)\beta + \mathbf{z}_i^T(t)\mathbf{b}_i + \epsilon_i(t) \\
    h_i(t|M_i(t), \mathbf{w}_i) = h_0(t) \exp\{\gamma^T \mathbf{w}_i + \alpha m_i(t)\}.
\end{cases}
\]

With this formulation, the hazard at age \(t\) for the \(i\)-th individual, with a true longitudinal profile \(M_i(t)\) up to time \(t\), can be expressed as

\[
h_i(t|M_i(t), \mathbf{w}_i) = h_0(t) \exp\{\gamma^T \mathbf{w}_i + \alpha (\mathbf{x}_i^T(t)\beta + \mathbf{z}_i^T(t)\mathbf{b}_i)\}.\]

### 3.5 Joint Model estimation

The main estimation methods proposed for joint models are maximum likelihood (Henderson et al., 2000; Hsieh et al., 2006; Wulfsohn and Tsiatis, 1997) and the Bayesian approach using MCMC techniques (Brown and Ibrahim, 2003; Chi and Ibrahim, 2006; Wany and Taylor, 2001).

**Maximum likelihood approach**

Maximum likelihood estimation for joint models is one of the most traditional approaches and is based on the maximization of the log-likelihood corresponding to the joint distribution of the time-to-event and the longitudinal outcomes, that can be written as \(\{T_i, \delta_i, y_i\}\). Note that we have omitted covariates \(\mathbf{w}_i\) to ease the notation. In order to
define this distribution, we will assume that the vector of random effects $\mathbf{b}_i$ underlies both the survival and longitudinal processes (so, in this case, random effects account for both the correlation between measurements in the longitudinal outcome and the association between the longitudinal and event outcomes). In other words, we are assuming that survival times and longitudinal covariates are independent given the random effects. We can write

$$p(T_i, \delta_i, \mathbf{y}_i | \mathbf{b}_i; \theta) = p(T_i, \delta_i | \mathbf{b}_i; \theta) p(\mathbf{y}_i | \mathbf{b}_i; \theta)$$ (3.7)

$$p(\mathbf{y}_i | \mathbf{b}_i; \theta) = \prod_j p\{y_i(t_{ij}) | \mathbf{b}_i; \theta_y, \beta\}$$ (3.8)

where $\theta = (\theta^T_t, \theta^T_y, \theta^T_b)^T$ is the parameter vector, with $\theta_t$ denoting the parameters for the survival outcome, $\theta_y$ the parameters for the longitudinal outcome and $\theta_b$ the unique parameters of the random-effects covariance matrix. Moreover, $\mathbf{y}_i$ denotes the vector of the $n_i$ longitudinal outcomes for the $i$-th subject and $p(\cdot)$ is an appropriate probability density function.

Under the conditional independence assumptions and using (3.7) and (3.8), the joint log-likelihood contribution for the $i$-th subject can be formulated as

$$\log p(T_i, \delta_i, \mathbf{y}_i; \theta) = \log \int_{\mathbf{b}_i} p(T_i, \delta_i | \mathbf{b}_i; \theta_t, \beta) \prod_j p\{y_i(t_{ij}) | \mathbf{b}_i; \theta_y, \beta\} p(\mathbf{b}_i; \theta_b) d\mathbf{b}_i,$$ (3.9)

where $p\{y_i(t_{ij}) | \mathbf{b}_i; \theta_y, \beta\}$ is the univariate normal density for the longitudinal responses and $p(\mathbf{b}_i; \theta_b)$ is the multivariate normal density for the random effects and the likelihood of the survival part is written as

$$p(T_i, \delta_i | \mathbf{b}_i; \theta_t, \beta) = [h_i(T_i | \mathcal{M}_i(T_i); \theta_t, \beta)]^{\delta_i} S_i(T_i | \mathcal{M}_i(T_i); \theta_t, \beta) = [h_i(T_i | \mathcal{M}_i(T_i); \theta_t, \beta)]^{\delta_i} \exp \left\{- \int_0^t h_i(s | \mathcal{M}_i(s); \theta_t, \beta) ds\right\},$$ (3.10)

The maximization of the log-likelihood function (3.9) with respect to $\theta$ is a computationally challenging task mainly because the integrals in (3.9) and (3.10) do not have analytical solutions, except in very special cases. From the two integrals, the one with respect to the random effects is the main computational bottleneck. The integral in the definition of the survival can be relatively efficiently approximated using 7-point or 15-point Gauss-Kronrod rule. However, the integral in (3.9) with respect to the random effects becomes computationally demanding to approximate as its dimensionality increases (Rizopoulos, 2012). Standard numerical integration techniques such as Gaussian quadrature rules and Monte Carlo sampling have been successfully applied in the joint modeling framework but still remain relatively computationally demanding. For the maximization of the log-likelihood function $l(\theta) = \sum_i \log p(T_i, \delta_i, \mathbf{y}_i; \theta)$ with respect to $\theta$, the Expectation-Maximization (EM) algorithm has been traditionally used (Dempster et al., 1977).
Bayesian approach

Under the Bayesian approach, estimation of the parameters of the joint model is done using Markov chain Monte Carlo (MCMC) algorithms. In a similar way as in the likelihood estimation we have

\[
p(T_i, \delta_i, y_i | b_i; \theta) = p(T_i, \delta_i | b_i; \theta) p(y_i | b_i; \theta)
\]

\[
p(y_i | b_i; \theta) = \prod_j p(y_i(t_{ij}) | b_i; \theta).
\]

Under the assumptions we have already presented, the posterior distribution is analogous to

\[
p(\theta, b|T_i, \delta_i, y_i) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} p(y_i(t_{ij}) | b_i, \theta) p(T_i, \delta_i | b_i, \theta) p(b_i) p(\theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{n_i} p(y_i(t_{ij}) | b_i, \theta) p(T_i, \delta_i | b_i, \theta) p(b_i) p(\theta),
\]

where

\[
p(y_i(t_{ij}) | b_i, \theta) = \exp \left\{ \frac{[y_i(t_{ij}) \psi_{ij}(b_i) - c \{ \psi_{ij}(b_i) \}]}{a(\varphi)} - d(y_i(t_{ij}), \varphi) \right\},
\]

with \( \psi_{ij}(b_i) \) and \( \varphi \) denoting the natural and dispersion parameters in the exponential family, respectively, \( c(\cdot), a(\cdot) \) and \( d(\cdot) \) being known functions specifying the member of the exponential family. For the survival part we have

\[
p(T_i, \delta_i | b_i, \theta) = [h_i(T_i | \mathcal{M}_i(T_i))]^{\delta_i} S_i(T_i | \mathcal{M}_i(T_i))
\]

\[
= [h_i(T_i | \mathcal{M}_i(T_i))]^{\delta_i} \exp \left\{ - \int_0^{T_i} h_i(s | \mathcal{M}_i(s)) \, ds \right\},
\]

and the integral in the definition of the survival function in (3.13)

\[
S_i(t | \mathcal{M}_i(t), \theta) = \exp \left\{ - \int_0^t h_0(s) \exp [\alpha m_i(t)] \, ds \right\}
\]

does not have a closed-form solution and, hence, a numerical method should be employed for its computation.

3.6 Joint Model illustration: TIBET

In this section we are going to apply the joint modeling techniques to a particular case study, the TIBET clinical trial. In the TIBET study there are two observed longitudinal outcomes for each patient at time \( t \), CD4 cell counts and viral load. As stated before,
we are interested in the survival outcome and we wish to account for the effect of a longitudinal variable as a time-dependent covariate. Thus, we are going to fit joint models separately for time until treatment reinitiation ($T_1$) and time until treatment interruption ($T_1 + T_2$). Since we can not include two longitudinal variables in the fitting of a joint model through packages JM and JMbayes, we have fit joint models separately for CD4 cell counts and for viral load. When fitting joint models, we use $\sqrt{CD4}$ and the logarithmic transformation of viral load to obtain linearity and to have a similar range of values in both variables. A choice that we have to make before fitting the models is the approximation of the log baseline hazard function. We have chosen to approximate it with splines, because it is the only common option in both packages and it allows to compare the results.

3.6.1 Models for time to treatment reinitiation $T_1$

We proceed by fitting joint models to study the time until treatment reinitiation with information from the longitudinal variables.

To study the time to reinitiation ($T_1$) adjusting for the longitudinal variable $\sqrt{CD4}$, we have fitted a joint model using as a baseline covariate for the survival part the information on the logarithm of the viral load that the patient had before entering the study ($\log(VL_{pret})$). For the longitudinal submodel, we are assuming a model with random intercept and random slope. The following joint model was fitted:

\[
\begin{align*}
\sqrt{CD4}_i(t) &= m_i(t) + \varepsilon_i(t) = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t + \epsilon_i(t) \\
\lambda_i(t|\mathcal{M}_i(t), w_i) &= h_0(t) \exp\{\gamma \log(VL_{pret}) + \alpha m_i(t)\},
\end{align*}
\]

where $b_i \sim \mathcal{N}_2(0, D)$ and $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2_{\epsilon})$ with $D$ being an unstructured $2 \times 2$ matrix for the random effects.

Parameter estimates can be found in Table 3.1 along with their 95% confidence intervals (for jointModel()) and credibility intervals (for jointModelBayes()). We can see that estimations from both packages are quite similar. In this regard, both show no correlation between the random effects (confidence and credibility intervals for $\rho_{b0b1}$ both contain zero) and a negative value for the association parameter $\alpha$. Thus, the CD4 cell counts (expressed as $\sqrt{CD4}$) have a slightly negative association with the risk of reinitiating therapy, a result that points in the direction of what we have already explained, that high levels of CD4 lengthen the time until treatment reinitiation.
Table 3.1: Joint models for time to reinitiation ($T_1$) adjusting for the longitudinal variable $\sqrt{\text{CD4}}$: parameter estimates, confidence and credibility intervals and AICs and DICs

<table>
<thead>
<tr>
<th>Parameters</th>
<th>JM</th>
<th>95% CI</th>
<th>JMbayes</th>
<th>95% Cred. Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Longitudinal submodel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>26.67</td>
<td>(25.91, 27.43)</td>
<td>27.38</td>
<td>(26.85, 27.84)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.12</td>
<td>(-0.13, -0.11)</td>
<td>-0.15</td>
<td>(-0.17, -0.12)</td>
</tr>
<tr>
<td>$\sigma_{b_0}$</td>
<td>4.11</td>
<td>(3.50, 4.72)</td>
<td>4.33</td>
<td>(3.71, 5.02)</td>
</tr>
<tr>
<td>$\sigma_{b_1}$</td>
<td>0.09</td>
<td>(0.09, 0.09)</td>
<td>2.08</td>
<td>(1.81, 2.41)</td>
</tr>
<tr>
<td>$\rho_{b_0b_1}$</td>
<td>0.11</td>
<td>(-0.13, 0.44)</td>
<td>-0.12</td>
<td>(-0.26, 0.11)</td>
</tr>
<tr>
<td>$\sigma_{\epsilon}$</td>
<td>2.3</td>
<td>(2.21, 2.39)</td>
<td>2.16</td>
<td>(2.08, 2.24)</td>
</tr>
<tr>
<td><strong>Survival submodel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.55</td>
<td>(0.17, 0.93)</td>
<td>1.01</td>
<td>(0.75, 1.29)</td>
</tr>
<tr>
<td><strong>Association</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>-0.29</td>
<td>(-0.38, -0.19)</td>
<td>-0.33</td>
<td>(-0.39, -0.27)</td>
</tr>
<tr>
<td><strong>Goodness of fit</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC / DIC</td>
<td>8198</td>
<td></td>
<td>8479</td>
<td></td>
</tr>
</tbody>
</table>

Restricted maximum likelihood estimates in the JM case. AIC represents the Akaike information criterion, DIC stands for the deviance information criterion, CI stands for confidence interval and Cred. Int. means credibility interval. The survival model considers log(VLpret) as a covariate.

To study the time to reinitation ($T_1$) adjusting for the longitudinal variable log (VL), we have fitted a joint model including in the survival submodel the information given by CD4 nadir ($\sqrt{\text{CD4N}}$), which is the lowest level of CD4 that subjects have ever had. In this scenario, we have chosen to include only random intercept in the survival model because when including also the random slope, the Bayesian approach did not converge for $\sigma_{b_1}$ and $\rho_{b_0b_1}$. Therefore, the following joint model was fitted:

$$
\begin{align*}
\log V_{Li}(t) &= m_i(t) + \epsilon_i(t) = (\beta_0 + b_{0i}) + \beta_1 t + \epsilon_i(t) \\
h_i(t|M_i(t), w_i) &= h_0(t) \exp\{\gamma \sqrt{\text{CD4N}} + \alpha m_i(t)\},
\end{align*}
$$

(3.16)

where $b_i \sim \mathcal{N}(0, \sigma_{b_0}^2)$ and $\epsilon_i(t) \sim \mathcal{N}(0, \sigma_{\epsilon}^2)$.

Estimation of the parameters and confidence and credibility intervals can be found in Table 3.2. We observe that although estimates for the parameters in the longitudinal and survival submodels are almost the same for packages JM and JMbayes, this is not the case for the estimate for the association between both processes. Whereas the frequentist approach reports a non significant association between longitudinal and survival processes, the Bayesian one acknowledges this link (although it reports an estimate close to zero). This can be due to a lack of converge in the Bayesian approach and running more iterations might lead to an estimate closer to the one in the frequentist approach.
### Table 3.2: Joint models for time to reinitiation ($T_1$) adjusting for the longitudinal variable log (VL): parameter estimates, confidence and credibility intervals and AICs and DICs

<table>
<thead>
<tr>
<th>Parameters</th>
<th>JM</th>
<th>JMbayes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Longitudinal submodel</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>3.88 (3.76, 4.007)</td>
<td>3.89 (3.77, 4.02)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.006 (0.005, 0.007)</td>
<td>0.006 (0.005, 0.008)</td>
</tr>
<tr>
<td>$\sigma_{b_0}$</td>
<td>0.51 (0.34, 0.68)</td>
<td>0.56 (0.48, 0.63)</td>
</tr>
<tr>
<td>$\sigma_{\varepsilon}$</td>
<td>0.79 (0.76, 0.82)</td>
<td>0.79 (0.76, 0.82)</td>
</tr>
<tr>
<td><strong>Survival submodel</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-0.13 (-0.17, -0.08)</td>
<td>-0.13 (-0.17, -0.10)</td>
</tr>
<tr>
<td><strong>Association</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.49 (-0.21, 1.18)</td>
<td>0.29 (0.06, 0.50)</td>
</tr>
<tr>
<td><strong>Goodness of fit</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC / DIC</td>
<td>4515</td>
<td>4592</td>
</tr>
</tbody>
</table>

Restricted maximum likelihood estimates in the JM case. AIC represents the Akaike information criterion, DIC stands for the deviance information criterion, CI stands for confidence interval and Cred. Int means credibility interval. The survival model considers $\sqrt{\text{CD4N}}$ as a covariate.

#### 3.6.2 Models for time to treatment interruption $T_1 + T_2$

In the same way as we did for time to treatment reinitiation, we are going to fit models for the time until treatment interruption with information from the longitudinal covariates.

The models are basically the same ones as fitted for $T_1$, but keeping in mind that we are now modeling the time from the first treatment interruption until the second.

The results for the time to withdrawal adjusting for the longitudinal $\sqrt{\text{CD4}}$ using the level of viral load pretherapy as a baseline covariate for the survival part can be found in Table 3.3. We can see that the coefficient for the baseline covariate log (VLpret) is non significant in both estimations. For the coefficient correlation between random effects, both functions report intervals containing zero suggesting independent random effects.

The packages give very different estimations for $\sigma_{b_1}$. It is plausible that, although package JM gives an estimation, in reality the coefficient is non identifiable and for this reason JMbayes does not converge. The Bayesian estimation reports an association coefficient non significant whereas in the frequentist one it is significant (but very close to zero). It is important to note that the model does not capture the relation between time to withdrawal and the longitudinal CD4 cell counts. This can be due to either the modeling of the commulative time $T_1 + T_2$ until treatment interruption instead of the depending censored gap time $T_2$ or the limitation of not being able to include both longitudinal covariates at the same time.
Table 3.3: Joint models for time to withdrawal ($T_1 + T_2$) adjusting for the longitudinal variable $\sqrt{CD4}$: parameter estimates, confidence and credibility intervals and AICs and DICs

<table>
<thead>
<tr>
<th>Parameters</th>
<th>JM</th>
<th>95% CI</th>
<th>JMbayes</th>
<th>95% Cred. Int</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Longitudinal submodel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>24.08</td>
<td>(23.16, 24.98)</td>
<td>24.18</td>
<td>(23.46, 24.89)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.0002</td>
<td>(-0.01, 0.01)</td>
<td>-0.02</td>
<td>(-0.06, 0.02)</td>
</tr>
<tr>
<td>$\sigma_{b_0}$</td>
<td>3.66</td>
<td>(2.97, 4.35)</td>
<td>3.64</td>
<td>(2.98, 4.35)</td>
</tr>
<tr>
<td>$\sigma_{b_1}$</td>
<td>0.04</td>
<td>(0.04, 0.04)</td>
<td>4.17</td>
<td>(3.51, 4.95)</td>
</tr>
<tr>
<td>$\rho_{b_0b_1}$</td>
<td>-0.32</td>
<td>(-0.62, 0.074)</td>
<td>0.0002</td>
<td>(-0.17, 0.37)</td>
</tr>
<tr>
<td>$\sigma_{\varepsilon}$</td>
<td>2.73</td>
<td>(2.62, 2.84)</td>
<td>2.72</td>
<td>(2.61, 2.84)</td>
</tr>
<tr>
<td><strong>Survival submodel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.32</td>
<td>(-0.07, 0.72)</td>
<td>-0.19</td>
<td>(-0.54, 0.14)</td>
</tr>
<tr>
<td><strong>Association</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.08</td>
<td>(0.007, 0.16)</td>
<td>0.0006</td>
<td>(-0.09, 0.08)</td>
</tr>
<tr>
<td><strong>Goodness of fit</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC / DIC</td>
<td>7004</td>
<td></td>
<td>7530</td>
<td></td>
</tr>
</tbody>
</table>

Restricted maximum likelihood estimates in the JM case. AIC represents the Akaike information criterion, DIC stands for the deviance information criterion, CI stands for confidence interval and Cred. Int means credibility interval. The survival model considers log(VLpret) as a covariate.

To model time to withdrawal adjusting for the longitudinal variable log (VL), we have fitted the same joint model as when we were modeling time to reinitiation.

Results for the parameter estimates and its confidence and credibility intervals can be found in Table 3.4. Both of them produce a similar fit. The association parameter is negative in both cases ($\alpha = -0.87$ for JM and $\alpha = -1.17$ for JMbayes), showing that an increase in the viral load decreases the risk to withdrawal, i.e. if the viral load of an individual increases, its time until treatment withdrawal will be longer.

In all these models we could only include one of the longitudinal variables and, thus, the information about the relationship between them could not be taken into account. Therefore, there is the necessity to develop a methodology to include more than one longitudinal variable in the predictor of the survival model. Moreover, because sequential survival times can be related (and they often are) it is also of importance to include them together in the model and use the information of their correlation. In the next Chapter we develop a methodology that allows to include in the joint model information on more than one longitudinal variable and more than one sequential survival time.
Table 3.4: Joint models for time to withdrawal ($T_1 + T_2$) adjusting for the longitudinal variable log (VL): parameter estimates, confidence and credibility intervals and AICs and DICs

<table>
<thead>
<tr>
<th>Parameters</th>
<th>JM</th>
<th>JMbayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>95% CI</td>
</tr>
<tr>
<td><strong>Longitudinal submodel</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>3.81</td>
<td>(3.65, 3.98)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.02</td>
<td>(-0.02, -0.016)</td>
</tr>
<tr>
<td>$\sigma_{b0}$</td>
<td>0.55</td>
<td>(0.32, 0.78)</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>1.14</td>
<td>(1.10, 1.19)</td>
</tr>
<tr>
<td><strong>Survival submodel</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-0.07</td>
<td>(-0.13, -0.005)</td>
</tr>
<tr>
<td><strong>Association</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>-0.87</td>
<td>(-1.51, -0.24)</td>
</tr>
<tr>
<td><strong>Goodness of fit</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC / DIC</td>
<td>4575</td>
<td></td>
</tr>
</tbody>
</table>

Restricted maximum likelihood estimates in the JM case. AIC represents the Akaike information criterion, DIC stands for the deviance information criterion, CI stands for confidence interval and Cred. Int means credibility interval. The survival model considers $\sqrt{CD4N}$ as a covariate.
Chapter 4

Methodology JMseq\((K, L)\)

4.1 Introduction

The majority of the joint models developed so far in the statistical literature focus on univariate time-to-event data. Chi and Ibrahim (2006) presented a model for multivariate longitudinal and multivariate survival data (but in that case the survival times were parallel and not sequential as the ones we aim to study in this work) and Liu and Huang (2009) studied a repeated measures process and a recurrent event process subject to a dependent terminal event (for example, higher CD4 cell counts are associated with lower risk of recurrent opportunistic diseases, both correlated with mortality). However, and although there are many potential applications, there is not much literature on the jointly modelization of longitudinal data with several types of failure time random variables, such as time to cancer relapse at two different organs, time to cancer relapse and death, time to first and second infection or, such as in the TIBET case, time without treatment and time with treatment.

In this kind of studies, where two, or more, consecutive times are observed and the censoring acts on their sum, the estimation of the joint survival function of the first time \((T_1)\) and the second \((T_2)\) has to take into account that \(T_2\) is only observed if \(T_1 + T_2\) does not exceed the total time of follow-up. Thus, the censoring mechanism acting on \(T_2\) will depend on \(T_1\). Because these variables are typically correlated, it is of great interest to examine their joint association with longitudinal markers. Hence it is important to develop a multivariate survival model capable of linking relevant longitudinal markers to a model for multivariate survival data.

Our goal is to link the information of the \(L\) longitudinal markers and a model for multivariate survival data (assuming that we have \(K\) survival times). To do so, we proceed as follows. We begin by constructing the hazard for an individual with two survival times and one longitudinal variable and then extend the case to \(K\) survival times (again with only one longitudinal variable). Then, we focus on more than one longitudinal variable. First, we study the case of two longitudinal variables (with only one survival time) and then we construct the hazard for \(L\) longitudinal variables and one survival time. Last step is taking into account both more than one survival time and more than one longitudinal variable. To fix ideas, we start by studying the case where \(K = 2\) and \(L = 2\) (two survival times and two longitudinal variables). This might help us set the basis for reaching our goal, the model with \(K\) survival times and \(L\) longitudinal variables.
As we have mentioned, this work was motivated by the TIBET clinical trial in which we have to model two sequential times ($T_1$ and $T_2$ with $T_2$ being observable only after $T_1$ have been observed) with longitudinal measurements. Our purpose is to take the longitudinal measurements as a markers for the times to events. If there is an association between the longitudinal and the survival processes, a joint modeling must be fitted to have proper model estimations for both processes.

4.2 $K$ survival times and one longitudinal variable

Multivariate survival data arises when we have a sequence of events $\mathcal{E}_1,\ldots,\mathcal{E}_K$ and its failure times $T_1,\ldots,T_K$ collected from the same individual. These failure times can be repetitions of the same kind of event or events of different type. We can distinguish (Lin et al., 1999) between two different types of multivariate data:

1. **Parallel data** arises when for each individual we observe several failure times $(T_1,\ldots,T_K)$ that do not satisfy any order restrictions and, thus, they can be observed in parallel.

2. **Longitudinal data (or sequential data)** arises when for each individual we observe a sequence of events over time (similar events that may occur several times for each subject). The times can correspond to events of the same kind (recurrent events) or they can be transitions between different states.

For our purposes, we will focus on sequential data, where we observe a sequence of times $T_{1i},\ldots,T_{Ki}$ for the $i$-th individual. The times are observed in order (first $T_{1i}$, then $T_{2i}$ and so on) and they occur sequentially so, for each individual, $T_j$ can only be observed if $T_1,\ldots,T_{j-1}$ have been previously observed. Because of this, subsequent times can depend on their previous times. The TIBET clinical trial mentioned in Chapter 2 belongs to this case because its transitions occur sequentially.

4.2.1 $K = 2$, $L = 1$: JMseq(2, 1)

We will study the situation of two survival times $T_1^*$ and $T_2^*$ that may be right censored by the censoring times $C_1$ and $C_2$ respectively. So, for each individual $i = 1,\ldots,n$ and for each sequential time $k = 1,2$, instead of observing $T_{ki}^*$, we observe $(T_{ki},\delta_{ki})$ where $T_{ki} = \min (T_{ki}^*, C_{ki})$ and $\delta_{ki} = I (T_{ki}^* \leq C_{ki})$, the indicator of whether $T_{ki}^*$ is censored or not. Again, instead of the true values of the biomarker $m_i(t_{ij})$ we observe $y_i(t_{ij})$ for $i = 1,\ldots,n$ and $j = 1,\ldots,n_i$ and where $t_{ij}$ are the calendar times where the measurements are taken.

If $T_1^*$ and $T_2^*$ are the two survival times, we can define the joint survival function as

$$S(t_1,t_2) = P (T_1^* > t_1, T_2^* > t_2)$$

and the distribution function can be written as

$$F(t_1,t_2) = P (T_1^* \leq t_1, T_2^* \leq t_2).$$
From (4.1) we can formulate the marginal survival functions as

\[ S_k(t_k) = P(T_K^* > t_k) \]

and so we will have \( S_1(t_1) = S(t_1, 0) \) and \( S_2(t_2) = S(0, t_2) \).

We can proceed in a similar way with (4.2) and the marginal distribution functions will be \( F_1(t_1) = F(t_1, 0) \) and \( F_2(t_2) = F(0, t_2) \).

Because the events occur to the same individual, the times will not be, in general, independent. So the PH Cox model cannot be used in this case because it assumes conditional independence of the failure times given the covariates.

For our sequence of times, where \( T_2 \) can not be observed until \( T_1 \) has been observed, it is neither possible to use simple pseudo-likelihood methods because they assume times are independent given the covariates. In this case, models can be formulated as a sequence of conditional distributions

For two sequential times \( T_1 \) and \( T_2 \), the sequence of conditional distributions can be expressed as

\[
F_1(t|w_i; \theta_1) = P(T_{1i} \leq t|w_i; \theta_1),
\]

\[
F_2|t_{1i}, (t|w_i, t_{1i}; \theta_2, \delta) = P(T_{2i} \leq t|w_i, t_{1i}; \theta_2, \delta),
\]

where \( \delta \) is the parameter association between \( T_1 \) and \( T_2 \), \( \theta_1 \) and \( \theta_2 \) are the vectors of the parameters associated to the covariates of \( T_1 \) and \( T_2 \) models and \( t_{1i} \) is the observed time (censored or not) for \( T_1 \).

Considering two times to event, the data observed can be one of the three types:

1. \( T_{1i}^* \) is not observed. So, consequently, \( T_{2i}^* \) neither. As we do not have any information on \( T_{2i}^* \), we defined \( \delta_{2i} = 0 \). In this situation we have \( T_{1i} > C_{1i} \) and \( \delta_{1i} = \delta_{2i} = 0 \). Thus, the information that we gathered is \( (T_{1i}, 0) \) with \( T_{1i} = C_{1i} \) for the first time and \( (0, 0) \) for the second (because we have no information for \( T_{2i} \)).

2. \( T_{1i}^* \) is observed and \( T_{2i}^* \) is not observed. In this case, the information we have is that \( T_{1i} = T_{1i}^* \), \( T_{2i} > C_{2i} \) and, for the censoring indicator, \( \delta_{1i} = 1 \) and \( \delta_{2i} = 0 \). The information for the first time will be \( (T_{1i}, 1) \) with \( T_{1i} = T_{1i}^* \) and \( (T_{2i}, 0) \) with \( T_{2i} = C_{2i} \) for the second.

3. The situation in which both \( T_{1i}^* \) and \( T_{2i}^* \) are observed. In this case, \( T_{1i} = T_{1i}^* \), \( T_{2i} = T_{2i}^* \) and \( \delta_{1i} = \delta_{2i} = 1 \) and thus, the information that we have is \( (T_{1i}, 1) \) with \( T_{1i} = T_{1i}^* \) for the first time and \( (T_{2i}, 1) \) with \( T_{2i} = T_{2i}^* \) for the second.

In Figure 4.1 we can see data from three individuals together with the censoring indicators in each case. First subject belongs to the first type above described. It has been censored for the first survival time and, consequently, there is no information on its second time. The second subject has been observed (black point) for the first survival time and censored (white point) for the second. Last, subject number three has been observed for both first and second survival times.
Chapter 4. Methodology JMseq(\(K, L\))

\[ \delta_{11} = \delta_{21} = 0 \]

\[ \delta_{12} = 1, \ \delta_{22} = 0 \]

\[ \delta_{13} = \delta_{23} = 1 \]

Figure 4.1: Types of observed data. The black points represent observations of the event of interest. The white ones stand for censored observations.

Based on these possibilities for observed data, the likelihood function may be written as

\[
L(\theta_1, \theta_2, \delta) = \prod_{i=1}^{n} \left[ f_1(t_{1,i}|w_i; \theta_1)^{\delta_{1i}} S_1(t_{1,i}|w_i; \theta_1)^{(1-\delta_{1i})} \right] \\
\left[ f_2(t_{2,i}|w_i, t_{1,i}; \theta_2, \delta)^{\delta_{2i}} S_2(t_{2,i}|w_i, t_{1,i}; \theta_2, \delta)^{(1-\delta_{2i})} \right]^{\delta_{1i}}. \quad (4.5)
\]

Thus, a model to analyze this kind of problems will be a model for \(T_1\) given baseline covariates and longitudinal information and a model for \(T_2\) given baseline covariates, longitudinal information and \(T_1\). First of all, we have to model the evolution of the longitudinal variables along time. We can assume that these variables follow different models such as linear in pieces, monotone (decreasing or increasing) with linear or parabolic trend and combinations of them (a two piecewise model until the first time to event and with a parabolic trend from then on, for instance). In this work, we assume our longitudinal variable to be linear in two pieces and changing in \(T_1\). Thus, in the end, we will have a value for the intercept and the slope for the first interval (from time zero to \(T_1\)) and a slope for the second interval (from \(T_1\) to \(T_1 + T_2\)). In the second interval, the intercept will be the current value at \(T_1\) for each individual. This assumption of a linear variable changing in \(T_1\) is due to the fact that because of the association between the longitudinal and survival processes, it is highly likely that the longitudinal trend changes because of the occurrence of the first event.
Given a two piecewise linear mixed model, the longitudinal variable can be written as
\[ y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) \left[ I_2(t_{ij}) t_{1,i} + (1 - I_2(t_{ij})) t_{ij} \right] \\
+ I_2(t_{ij}) (\beta_2 + b_{i2}) (t_{ij} - t_{1,i}) + \epsilon_i (t_{ij}), \quad (4.6) \]
where \( t_{1,i} \) is, as before, the observed time for \( T_1 \) and \( I_2(t_{ij}) \) is the indicator of whether the individual is at risk for the second event and can be written as \( I_2(t_{ij}) = I(t_{ij} \geq t_{1,i}) \). Thus, at a calendar time \( t_{ij} \) if the individual has not experienced \( E_1 \) yet, the longitudinal variable is modeled through an intercept \( (\beta_0) \) and an slope \( (\beta_1) \) and their random effects, whereas if \( t_{1,i} \) has been already observed, (4.6) includes \( \beta_2 \) to model the evolution of the variable in the second interval of time as well as the random effect for the \( i \)-th subject in this second interval. In this case, the amount of time the \( i \)-th individual has spent at this second interval or, in other words, the time the individual has been at risk for \( T_2 \), is written as \( t_{ij} - t_{1,i} \).

From (4.6) we can see that we are assuming that both the intercept and the slope have random effects in the two sections. Otherwise, the model can be easily adjusted to the situation where only the intercept or the slope (or neither) have random effects. From now on we will assume random effects in both the intercept and the slope. As we have mentioned before, from (4.6) we can see that the model when the subject is at risk for \( T_2 \) takes information from the previous longitudinal trajectory. In other words, the intercept in this second interval is the current value for the subject at the moment in which the previous event of interest has occurred.

As we have already said, in this work we are going to focus on a model with linear trend changing in \( T_1 \). For that case the hazard for an individual that has not experimented \( E_1 \) yet can be written as
\[ h_{1,i}(t|M_i(t), w_i) = h_{1,0}(t) \exp \left\{ \gamma_1^T w_i + \alpha_{1,1} [(\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t] \right\}, \quad (4.7) \]
with \( \gamma_1 \) being the vector of coefficients of the baseline covariates \( w_i \) that have effect in \( T_1 \). In (4.7) we introduced the parameter \( \alpha_{k,i} \) that denotes the association between the the \( k \)-th sequential time and the \( l \)-th longitudinal variable (here, \( \alpha_{1,1} \) that denotes the association between the first survival time and the first (and only) longitudinal variable).

Thus, although it would have been easier to use \( \alpha_1 \) in (4.7), because there is only one longitudinal variable, we have chosen to use the more complex notation \( \alpha_{1,1} \) in order to be consistent with next sections. Moreover, note that in (4.7) \( t \) stands for the time that the individual is at risk for \( T_1 \).

In a quite similar way and introducing \( \delta_{2,1} \) to describe the relation between \( T_1 \) and \( T_2 \), the hazard for and individual that has experimented \( E_1 \) can be expressed as
\[ h_{2,i}(t|M_i(t_{1,i} + t), w_i, t_{1,i}) = h_{2,0}(t) \exp \left\{ \gamma_2^T w_i + \delta_{2,1} t_{1,i} \right\} \\
+ \alpha_{2,1} [(\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t_{1,i} + (\beta_2 + b_{i2}) t], \quad (4.8) \]
where \( \gamma_2 \) is the vector of the coefficients of the baseline covariates affecting on the second interval and \( \alpha_{2,1} \) is the association between survival and longitudinal processes in that second interval. In (4.8) \( t \) represents the time the individual is at risk for \( T_2 \).

Note that whereas in (4.7) we can only use the information from \( M_i(t) \), in (4.8) we can use \( M_i(t_{1,i} + t) \) because the subject has already experienced \( E_1 \) and we can use the information until \( t_{1,i} + t \) to model the longitudinal variable in the second interval.
4.2.2 $L = 1$: JMseq($K, 1$)

In this section the notations and concepts for two sequential times will be extended for the case of $K$ survival times. Suppose we have $K$ survival times $T_k^*$ that may be right censored with censure $C_k$ for $k = 1, \ldots, K$. As before, for each individual we observe $(T_{ki}, \delta_{ki})$ where $T_{ki} = \min(T_k^*, C_k)$ and $\delta_{ki} = I(T_k^* \leq C_k)$ for $k = 1, \ldots, K$ and $i = 1, \ldots, n$.

If $T_1^*, \ldots, T_K^*$ are the $K$ survival times associated to an individual, we can define their joint survival function as

$$S(t_1, \ldots, t_K) = P(T_1^* > t_1, \ldots, T_K^* > t_K).$$

(4.9)

The distribution function can be written as

$$F(t_1, \ldots, t_K) = P(T_1^* \leq t_1, \ldots, T_K^* \leq t_K).$$

(4.10)

Following the notation, the marginals for the survival function are

$$S_1(t_1) = S(t_1, 0, \ldots, 0), \ldots, S_k(t_K) = S(0, \ldots, 0, t_K).$$

(4.11)

As in the case of two sequential times, PH Cox model cannot be used in this case because it assumes conditional independence of the failures given the covariates and neither can be applied simple pseudo-likelihood methods because they assume times are independent given the covariates and this is not the case for our sequence of times, where $T_j$ can not be observed until $T_1, \ldots, T_{j-1}$ have been observed.

In order to arrive to a general formulation, we denote

$$\Delta_k = \begin{pmatrix} \delta_{k,1} \\ \vdots \\ \delta_{k,k-1} \end{pmatrix}$$

(4.12)

as the vector of the relations for $T_k$ and the previous times to event $T_1, \ldots, T_{k-1}$. The components of this vector are expected to be decreasing in norm if $T_k$ is more influenced by $T_{k'}$ as $k'$ is closer to $k$. Hence, as we move forward in time the first sequential times tend to have less influence in the current time.

To ease the notation we also denote

$$T_{ki} = \begin{pmatrix} t_{1,i} \\ \vdots \\ t_{k-1,i} \end{pmatrix}$$

(4.13)

as the observed times at which the $i$-th individual has experienced the previous $k - 1$ events of interest.
With this notation, models for $T_1, T_2, \ldots, T_K$ given covariates $w_i$ can be written as

$$F_{k|T_{ki}}(t|w_i, T_{ki}; \theta_k, \Delta_{k}) = P(T_{ki} \leq t|w_i, T_{ki}; \theta_k, \Delta_{k}),$$

(4.14)

where, with $T_{ki}$ and $\Delta_{k}$ we are using the information we have gathered in the previous times.

As before, in order to construct a model capable of linking longitudinal variables to a model for multivariate survival data first of all we have to model the evolution of the longitudinal variables along time. We assume that we have a longitudinal variable that has a linear trend changing in every $T_k$. We denote, for $k \geq 2$, $I_k(t_{ij}) = I(t_{ij} \geq \sum_{p=1}^{k-1} t_{p,i})$ as the indicator of whether the $i$-th individual is at risk for $T_k$ at time $t_{ij}$ or, in other words, if $t_{ij}$ is bigger than the sum of the previous $k-1$ times. For this indicator, we define $I_1(\cdot) \equiv 1$ and $I_{K+1}(\cdot) \equiv 0$, meaning that $t_{ij}$ is always at risk at the beginning and never at risk after the last event. With this notation, the longitudinal variable for $K$ sequential survival times at a certain calendar time $t_{ij}$ can be written as

$$y_{ij} = (\beta_0 + b_{0i}) + \sum_{k=1}^{K} \left\{ I_k(t_{ij}) (\beta_k + b_{ki}) \left[ I_{k+1}(t_{ij}) t_{k,i} + (1 - I_{k+1}(t_{ij})) (t_{ij} - \sum_{p=1}^{k-1} t_{p,i}) \right] \right\} + \epsilon_i(t_{ij}),$$

(4.15)

where the indicator $I_k$ indicates if $t_{ij}$ is at risk (at least) for $T_{k-1}$ (and, thus, we have to take into account the effect of $\beta_k$ and $b_{ki}$) and the purpose of $I_{k+1}$ is to know if we are at risk for $T_{k+1}$ or not. With this two indicators, we can always know in which interval $t_{ij}$ is and therefore, calculated the value of the longitudinal variable according to it.

In Figure 4.2 we can see the construction of the longitudinal variable when the individual is at risk for a particular $T_k$. Apart from the intercept effect (fixed and linear, because we are considering a specific subject $i$ and thus we have to consider its deviation from the marginal), as $t_{ij}$ is bigger than $t_{p,i}$, for $p = 1, \ldots, k-1$, we have to include the effect of all previous times. Therefore, for each time $t_{p,i}$, meaning the time the individual has been at risk for $T_p$, we add its contribution $(\beta_p + b_{pi}) t_{p,i}$ to the longitudinal response. For the $k$-th interval of time, we have to add up the effect of the time the individual has been at risk for $T_k$, that is $t_{ij} - \sum_{k=1}^{p-1} t_{k,i}$, through $\beta_k$ and $b_{ki}$.

Therefore, in particular, if we know that $t_{ij}$ is at risk for a certain $T_k$ the variable can be written as follows

$$y_{ij} = (\beta_0 + b_{0i}) + \sum_{p=1}^{k-1} [(\beta_p + b_{pi}) t_{p,i}] + (\beta_k + b_{ki}) \left( t_{ij} - \sum_{p=1}^{k-1} t_{p,i} \right) + \epsilon_i(t_{ij}),$$

(4.16)

where $t_{ij} - \sum_{p=1}^{k-1} t_{p,i}$ is the time the individual has been at risk for $T_k$ or, in other words, how much larger is $t_{ij}$ from $\sum_{p=1}^{k-1} t_{p,i}$.
Chapter 4. Methodology JMseq(K,L) 30

Figure 4.2: Construction of the longitudinal variable with sequential times. For each interval of time we can see its contribution to the overall expression of the longitudinal variable.

Using what we have developed for the case of two sequential times, the hazard for an individual that has not yet arrived to \( T_1 \) is (4.7) and for an individual that has experimented \( E_1 \) but not yet \( E_2 \) is (4.8).

Analogously, for an individual that has experienced the first and second events and is \( t \) units at risk for \( T_3 \), the hazard can be written as

\[
\begin{align*}
 h_{3,i}(t \mid \mathcal{M}_i (t_{1,i} + t_{2,i} + t), w_i, t_{1,i}, t_{2,i}) &= h_{3,0}(t) \exp \left\{ \gamma_3 w_i + \delta_{3,1} t_{1,i} + \delta_{3,2} t_{2,i} \right. \\
 &\quad + \left. \alpha_{3,1} \left[ \beta_0 + b_{i0} + (\beta_1 + b_{i1}) t_{1,i} + (\beta_2 + b_{i2}) t_{2,i} + (\beta_3 + b_{i3}) t \right] \right\}.
\end{align*}
\]  
(4.17)

With the new variables \( T_{ki} \) and \( \Delta_k \) that we have already introduced, we can extend the expressions for the hazards that we have formulated before to the general case where there are \( K \) sequential times. Thus the risk for a certain individual \( i \) at \( t \) units of risk for the \( k \)-th time will be

\[
\begin{align*}
 h_{k,i} \left( t \mid \mathcal{M}_i \left( \sum_{p=1}^{k-1} t_{p,i} + t \right), w_i, T_{k,i} \right) &= h_{k,0}(t) \exp \left\{ \gamma_k w_i + \Delta_k^T T_{ki} \right. \\
 &\quad + \left. \alpha_{k,1} \left[ (\beta_0 + b_{i0}) + \sum_{p=1}^{k-1} (\beta_p + b_{ip}) t_{p,i} + (\beta_k + b_{ik}) t \right] \right\}.
\end{align*}
\]  
(4.18)

4.3 One survival time and \( L \) longitudinal variables

As we have mentioned, the majority of the work in the joint modeling literature has focused on models with only one longitudinal outcome. However, in many longitudinal studies, patients are measured repeatedly for several outcomes assumed to be associated with the survival times and that are potentially predictive for the time-to-event. In this
Section we will introduce the methodology for the joint modeling of \(L\) longitudinal variables and one survival time, beginning with the case of two longitudinal variables and extending it to the general case of \(L\) variables.

### 4.3.1 \(K = 1, L = 2\): JMseq(1, 2)

As before, we will assume that our variables follow a linear piecewise model. Thus, the two longitudinal variables can be written as

\[
y_{1ij} = (\beta_{10} + b_{1i0}) + (\beta_{11} + b_{1i1})t_{ij} + \epsilon_{1i}(t_{ij}), \tag{4.19}
\]

\[
y_{2ij} = (\beta_{20} + b_{2i0}) + (\beta_{21} + b_{2i1})t_{ij} + \epsilon_{2i}(t_{ij}), \tag{4.20}
\]

where using the subindex 1 or 2, we specified that the \(\beta\)'s and the random effects are different for each variable.

And the hazard for a particular individual \(i\) will be

\[
h_{1,i}(t|M_i(t), w_i) = h_{1,0}(t) \exp \left\{ \gamma^T w_i + \alpha_{1,1} [(\beta_{10} + b_{1i0}) + (\beta_{11} + b_{1i1})t] + \alpha_{1,2} [(\beta_{20} + b_{2i0}) + (\beta_{21} + b_{2i1})t] \right\}, \tag{4.21}
\]

where for \(\alpha_{1,1}\) and \(\alpha_{1,2}\) we denote the association between the first survival time and the first and second longitudinal variables, respectively.

We denote \(b_{1i} = (b_{1i0}, b_{1i1})^T\) and \(b_{2i} = (b_{2i0}, b_{2i1})^T\) the vector for the random effects of the two variables and, in order to model the dependency between the two longitudinal variables in the case of one survival time, we denote \(\Gamma_{1,2}\) as its symmetric covariance matrix for the random effects and we can write

\[
\Gamma_{1,2} = \begin{pmatrix}
D_{11} & D_{12} \\
D_{21} & D_{22}
\end{pmatrix}, \tag{4.22}
\]

where

\[
D_{21} = D_{12}^T.
\]

To ease the notation for next sections, we define \(D_{lm}\) as the covariance matrix between the \(l\)-th and the \(m\)-th longitudinal variables and \(\sigma_{lm,pq}\) as the covariance between the \(p\)-th and the \(q\)-th random effect of the \(l\)-th and \(m\)-th longitudinal variables, respectively. Thus, (4.22) can be written more explicitly as

\[
\Gamma_{1,2} = \begin{pmatrix}
\sigma_{11,00} & \sigma_{11,01} & \sigma_{12,00} & \sigma_{12,01} \\
\sigma_{11,11} & \sigma_{12,10} & \sigma_{12,11} \\
\sigma_{22,00} & \sigma_{22,01} & \sigma_{22,11}
\end{pmatrix}.
\]

(4.23)
Therefore, $\Gamma_{1,2}$ is a symmetric matrix that has 4 rows and 4 columns and 10 different components. This matrix models the relation between the random effects for the two longitudinal variables. For example, if $\sigma_{12,00} = \text{cov}(b_{1i0}, b_{2i0}) > 0$, it means that the levels of $y_1$ and $y_2$ at the intercept level from their mean population profile move towards the same direction: if $b_{1i0}$ goes up, so does $b_{2i0}$. On the other hand, when $\text{cov}(b_{1i0}, b_{2i0}) < 0$, if $b_{1i0}$ goes up, $b_{2i0}$ will go down and the other way round.

### 4.3.2 $K = 1$: JMseq(1, L)

For $L$ longitudinal variables, and continuing with only one survival time, we can extend the notation that we already had in (4.19) and (4.20) for the case of two variables to the case of $L$ variables and construct a set of $L$ longitudinal variables as follows

\[
y_{1ij} = (\beta_{10} + b_{1i0}) + (\beta_{11} + b_{1i1}) t_{ij} + \epsilon_{1i} (t_{ij}), \tag{4.24}
\]

\[
y_{Lij} = (\beta_{L0} + b_{Li0}) + (\beta_{L1} + b_{Li1}) t_{ij} + \epsilon_{Li} (t_{ij}). \tag{4.25}
\]

If we denote $\alpha_{1,1}, \ldots, \alpha_{1,L}$ the associations between the survival time and the $L$ longitudinal variables and $\gamma_1$ the effect of the vector of baseline covariates $w$ for a certain individual $i$, the hazard for this individual can be written as

\[
h_{1,i}(t|M_i(t), w_i) = h_{1,0}(t) \exp \left\{ \gamma_1^T w_i + \alpha_{1,1} [(\beta_{10} + b_{1i0}) + (\beta_{11} + b_{1i1}) t] \\
+ \alpha_{1,2} [(\beta_{20} + b_{2i0}) + (\beta_{21} + b_{2i1}) t] + \ldots \\
+ \alpha_{1,L} [(\beta_{L0} + b_{Li0}) + (\beta_{L1} + b_{Li1}) t] \right\} \\
= h_{1,0}(t) \exp \left\{ \gamma_1^T w_i + \sum_{l=1}^{L} \alpha_{1,l} [(\beta_{l0} + b_{li0}) + (\beta_{l1} + b_{li1}) t] \right\}. \tag{4.26}
\]

As before, and using the notation introduced in the previous subsection, we can construct the block matrix $\Gamma_{1,L}$ that models the covariance between the random effects of all $L$ variables as

\[
\Gamma_{1,L} = \begin{pmatrix}
D_{11} & D_{12} & \cdots & D_{1L} \\
D_{21} & D_{22} & \cdots & D_{2L} \\
\vdots & \vdots & \ddots & \vdots \\
D_{L1} & D_{L2} & \cdots & D_{LL}
\end{pmatrix}
\]

The blocks of this block matrix are

- $L$ $2 \times 2$ matrices $D_{ll}$ with $l = 1, \ldots, L$ and each one with 3 different components.

- $\frac{L}{2}(L - 1)$ different $2 \times 2$ matrices $D_{lm}$ (because $D_{ml} = D_{lm}^T$) and each one with 4 different components.

To sum up, the matrix $\Gamma_{1,L}$ that gathers the information about the covariance between the random effects associated with the $L$ longitudinal variables is a $2L \times 2L$ symmetric matrix with $3L + 2L(L - 1)$ different components. In previous section, as we had 2
longitudinal variables, \( L = 2 \) and we had a \( 4 \times 4 \) matrix with \( 3 \cdot 2 + 2 \cdot 2 \cdot 1 = 10 \) different components. More information on the parameters and the dimension of \( \Gamma_{1,L} \) is available in Table 4.1 at the end of this Chapter.

4.4 \( K \) sequential times and \( L \) longitudinal variables

In this next step we aim to joint the information given by \( K \) sequential survival times with the information of \( L \) longitudinal variables to develop an expression for the hazard for an individual when we have more than one survival time and information on more than one longitudinal variable. First of all we study the case of 2 survival times and 2 longitudinal variables. Then, we extend it to the \( K \) sequential times and \( L \) longitudinal variables’ case.

4.4.1 \( K = 2, L = 2: \text{JMseq}(2,2) \)

If we have two different markers and each of them can be modeled as a longitudinal model over \( T_1 \) and \( T_2 \), we have the case of two longitudinal variables and two sequential times to event. The two longitudinal variables do not have to follow the same evolution thus, as before, first of all we have to choose a proper model for each longitudinal variable. As we choose to model them both with a piecewise model changing in \( T_1 \), the two variables can be written as in (4.6) but now we have to specify which variable (1 or 2) we are talking about. Thus, in this case, the variables can be written as

\[
y_{1ij} = (\beta_{10} + b_{1i0}) + (\beta_{11} + b_{1i1}) [I_2(t_{ij})t_{1,i} + (1 - I_2(t_{ij})) t_{ij}] \\
+ I_2(t_{ij}) (\beta_{12} + b_{1i2}) (t_{ij} - t_{1,i}) + \epsilon_i(t_{ij}), \tag{4.27}
\]

\[
y_{2ij} = (\beta_{20} + b_{2i0}) + (\beta_{21} + b_{2i1}) [I_2(t_{ij})t_{1,i} + (1 - I_2(t_{ij})) t_{ij}] \\
+ I_2(t_{ij}) (\beta_{22} + b_{2i2}) (t_{ij} - t_{1,i}) + \epsilon_i(t_{ij}), \tag{4.28}
\]

where, as in (4.6), we are denoting \( t_{1,i} \) as the time at which we have observed the first event for individual \( i \) and \( I_2(t_{ij}) = I(t_{ij} \geq t_{1,i}) \) as the indicator of whether \( t_{ij} \) is bigger than \( t_{1,i} \) or not. Thus, \( \beta_{k2} \) and \( b_{ki2} \) for \( k = 1,2 \) will have an effect on the hazard only if \( t_{ij} \) is bigger than \( t_{1,i} \) (that means, if at \( t_{ij} \) the individual is at risk for \( T_2 \)).

With the expression of the two variables we can construct the hazard for a particular individual \( i \) that has not yet experienced \( E_1 \) as follows

\[
h_{1,i}(t | \mathcal{M}_{i}(t), w_i) = h_{1,0}(t) \exp \left\{ \gamma^T w_i + \alpha_{1,1} [(\beta_{10} + b_{1i0}) + (\beta_{11} + b_{1i1}) t] \\
+ \alpha_{1,2} [(\beta_{20} + b_{2i0}) + (\beta_{21} + b_{2i1}) t] \right\}. \tag{4.29}
\]
And similarly, the hazard for an individual that has experienced $\mathcal{E}_1$ and therefore, is at risk for $T_2$, can be written as
\[
h_{2,i}(t|M_i(t_{1,i}+t),w_i,t_{1,i}) = h_{2,0}(t)\exp\left\{\gamma_2^T w_i + \delta_2 t_{1,i}
+ \alpha_{2,1} [(\beta_{10} + b_{1i0}) + (\beta_{11} + b_{1i1}) t_{1,i} + (\beta_{12} + b_{1i2}) t]
+ \alpha_{2,2} [(\beta_{20} + b_{2i0}) + (\beta_{21} + b_{2i1}) t_{1,i} + (\beta_{22} + b_{2i2}) t]\right\}.
\]

Note that (4.29) and (4.30) are similar to (4.7) and (4.8) but now we have to include the effect of the two longitudinal variables and, thus, we have two different association parameters for each hazard, $\alpha_{1,1}$ and $\alpha_{1,2}$ for the hazard related with the first event and $\alpha_{2,1}$ and $\alpha_{2,2}$ for the hazard related with the second.

In this scenario, we have more than one longitudinal variable and, therefore, we can construct the covariance matrix between their respective random effects. As we are considering two sequential times, each variable has three random effects ($b_{ki0}$, $b_{ki1}$ and $b_{ki2}$). Therefore, with the notation that we have already introduced, the covariance matrix for, for example, the random effects of the first variable can be written as
\[
D_{11} = \begin{pmatrix}
\sigma_{11,00} & \sigma_{11,01} & \sigma_{11,02} \\
\sigma_{11,10} & \sigma_{11,11} & \sigma_{11,12} \\
\sigma_{11,20} & \sigma_{11,21} & \sigma_{11,22}
\end{pmatrix}.
\]

Putting together the information obtained from the covariance matrix of each variable and the covariance matrix between the two variables, the covariance matrix that describes the relation between the random effects of both variables will be
\[
\Gamma_{2,2} = \begin{pmatrix}
D_{11} & D_{12} \\
D_{21} & D_{22}
\end{pmatrix},
\]

where
\[
D_{12} = \begin{pmatrix}
\text{cov}(b_{1i0},b_{2i0}) & \text{cov}(b_{1i0},b_{2i1}) & \text{cov}(b_{1i0},b_{2i2}) \\
\text{cov}(b_{1i1},b_{2i0}) & \text{cov}(b_{1i1},b_{2i1}) & \text{cov}(b_{1i1},b_{2i2}) \\
\text{cov}(b_{1i2},b_{2i0}) & \text{cov}(b_{1i2},b_{2i1}) & \text{cov}(b_{1i2},b_{2i2})
\end{pmatrix} = \begin{pmatrix}
\sigma_{12,00} & \sigma_{12,01} & \sigma_{12,02} \\
\sigma_{12,10} & \sigma_{12,11} & \sigma_{12,12} \\
\sigma_{12,20} & \sigma_{12,21} & \sigma_{12,22}
\end{pmatrix}.
\]

In conclusion,
\[
\Gamma_{2,2} = \begin{pmatrix}
\sigma_{11,00} & \sigma_{11,01} & \sigma_{11,02} & \sigma_{12,00} & \sigma_{12,01} & \sigma_{12,02} \\
\sigma_{11,10} & \sigma_{11,11} & \sigma_{11,12} & \sigma_{12,10} & \sigma_{12,11} & \sigma_{12,12} \\
\sigma_{11,20} & \sigma_{11,21} & \sigma_{11,22} & \sigma_{12,20} & \sigma_{12,21} & \sigma_{12,22}
\end{pmatrix}.
\]

Therefore, $\Gamma_{2,2}$ is a symmetric matrix that has 6 rows and 6 columns and 21 different components. For more information on the number of parameters or the dimension of the block matrices, see Table 4.1 at the end of this Chapter.
4.4.2 JMseq\((K, L)\)

The final step is to joint the information of \(K\) sequential survival times and \(L\) longitudinal variables. In a similar way as we did in (4.15), we can write the \(L\) longitudinal variables as

\[
y_{1ij} = (\beta_{10} + b_{1i0}) + \sum_{k=1}^{K} \{ I_k(t_{ij}) (\beta_{1k} + b_{1ik}) \left[I_{k+1}(t_{ij}) t_{k,i} + (1 - I_{k+1}(t_{ij})) (t_{ij} - \sum_{p=1}^{k-1} t_{p,i})\right]\} + \epsilon_{1i}(t_{ij}),
\]

\[(4.35)\]

\[
y_{2ij} = (\beta_{20} + b_{2i0}) + \sum_{k=1}^{K} \{ I_k(t_{ij}) (\beta_{2k} + b_{2ik}) \left[I_{k+1}(t_{ij}) t_{k,i} + (1 - I_{k+1}(t_{ij})) (t_{ij} - \sum_{p=1}^{k-1} t_{p,i})\right]\} + \epsilon_{2i}(t_{ij}),
\]

\[(4.36)\]

\[
\ldots
\]

\[
y_{Lij} = (\beta_{L0} + b_{Li0}) + \sum_{k=1}^{K} \{ I_k(t_{ij}) (\beta_{Lk} + b_{Lik}) \left[I_{k+1}(t_{ij}) t_{k,i} + (1 - I_{k+1}(t_{ij})) (t_{ij} - \sum_{p=1}^{k-1} t_{p,i})\right]\} + \epsilon_{3i}(t_{ij}),
\]

\[(4.37)\]

where each of them is thought as having a linear trend that changes in each survival time. This change at each \(T_k\) is modeled through the different \(K\) parameters \(\beta\).

Thus, in this situation, the hazard for an individual must include the information given by all the longitudinal variables, each one with its coefficient \(\alpha_{k,l}\) that denotes the association between the \(k\) event time and the \(l\)-th longitudinal variable. Thus, for a certain individual \(i\), the hazard for a particular time \(t\) between 0 and the first event can be written as

\[
h_{1,i}(t|\mathcal{M}_{i}(t), w_{i}) = h_{1,i0}(t) \exp \left\{ \gamma_1^T w_{i} + \alpha_{1,1} [ (\beta_{10} + b_{1i0}) + (\beta_{11} + b_{1i1}) t ] \\ + \alpha_{1,2} [ (\beta_{20} + b_{2i0}) + (\beta_{21} + b_{2i1}) t ] + \ldots \\
+ \alpha_{1,L} [ (\beta_{L0} + b_{L0}) + (\beta_{L1} + b_{Li1}) t ] \right\},
\]

\[(4.38)\]

where only the first two fixed and random effects have an effect on the hazard because as the individual has not experienced \(\mathcal{E}_1\) yet, we can only use information on the first interval and, thus, the longitudinal variable can be model with only the effects of the intercept (\(\beta_{0i}\) and \(b_{i0}\)) and the slope (\(\beta_{1i}\) and \(b_{i1}\)), for \(l = 1, \ldots, L\), in that first interval of time.

For the hazard for a time \(t\) between the first and the second event, we must include the effect of the first observed time \((t_{1,i})\) and the longitudinal variables expand to the second
interval of time. Therefore, as in (4.30) but including the $L$ longitudinal variables, the hazard for a time $t$ in the gap between the first and the second time-to-event will be

$$h_{2,i}(t|\mathcal{M}_i(t_{1,i} + t), \mathbf{w}_i, t_{1,i}) = h_{2,0}(t) \exp \left\{ \gamma_i^T \mathbf{w}_i + \delta_{2,i} t_{1,i} \right\} + \alpha_{2,1} [\beta_{10} + b_{1i0}] + \alpha_{2,2} [\beta_{20} + b_{2i0}] + \alpha_{2,L} [\beta_{L0} + b_{Li0}]$$

+ \alpha_{2,1} [(\beta_{11} + b_{1i1}) t_{1,i} + (\beta_{12} + b_{1i2}) t] + \alpha_{2,2} [(\beta_{21} + b_{2i1}) t_{1,i} + (\beta_{22} + b_{2i2}) t] + \ldots + \alpha_{2,L} [(\beta_{L1} + b_{Li1}) t_{1,i} + (\beta_{L2} + b_{L2i2}) t]

(4.39)

In general, for a general $k$, $k = 1, \ldots, K$, using variables $T_{ki}$ and $\Delta_k$ to include the effect of previous times and using the association variables $\alpha_{k,1}, \ldots, \alpha_{k,L}$ to use the information from the $L$ longitudinal variables, we can write the hazard for a particular individual $i$ that is at risk for the $k$-th event as

$$h_{k,i}(t|\mathcal{M}_i\left(\sum_{p=1}^{k-1} t_{p,i} + t\right), \mathbf{w}_i, T_{k,i}) = h_{k,0}(t) \exp \left\{ \gamma_k^T \mathbf{w}_i + \Delta_k^T T_{ki} \right\} + \alpha_{k,1} \left[ (\beta_{10} + b_{1i0}) + \sum_{p=1}^{k-1} (\beta_{1p} + b_{1ip}) t_{p,i} + (\beta_{1k} + b_{1ik}) t \right]$$

+ $\alpha_{k,2} \left[ (\beta_{20} + b_{2i0}) + \sum_{p=1}^{k-1} (\beta_{2p} + b_{2ip}) t_{p,i} + (\beta_{2k} + b_{2ik}) t \right]$

+ $\ldots$

+ $\alpha_{k,L} \left[ (\beta_{L0} + b_{Li0}) + \sum_{p=1}^{k-1} (\beta_{Lp} + b_{Lip}) t_{p,i} + (\beta_{Lk} + b_{Lik}) t \right]$

= $h_{k,0}(t) \exp \left\{ \gamma_k^T \mathbf{w}_i + \Delta_k^T T_{ki} \right\} + \sum_{l=1}^{L} \alpha_{k,l} \left[ (\beta_{l0} + b_{li0}) + \sum_{p=1}^{k-1} (\beta_{lp} + b_{lip}) t_{p,i} + (\beta_{lk} + b_{lik}) t \right]$

(4.40)

The design of the covariance matrix for the random effects of all variables is a bit more complicated in this case because we have to acknowledge not only the information about the $L$ longitudinal variables but also the information about the $K$ sequential survival times. The matrix design in blocks will be

$$\Gamma_{K,L} = \begin{pmatrix}
D_{11} & D_{12} & \cdots & D_{1L} \\
D_{21} & D_{22} & \cdots & D_{2L} \\
\vdots & \vdots & \ddots & \vdots \\
D_{L1} & D_{L2} & \cdots & D_{LL}
\end{pmatrix}.$$
And the blocks of this block matrix are

- \( L (K + 1) \times (K + 1) \) matrices \( D_i \) with \( i = 1, \ldots, L \), each one with \( \frac{1}{2} (K + 1) (K + 2) \) different components.

- \( \frac{1}{2} L (L - 1) \) different \( (K + 1) \times (K + 1) \) matrices \( D_{lm} \) (because \( D_{mt} = D_{lm}^T \)), each one with \( (K + 1) (K + 1) \) different components.

The construction of this matrices is similar to the one we did in the case of two sequential survival times and two longitudinal variables. More information on parameters and dimensions is available in Table 4.1 at the end of this Chapter.
Table 4.1: Dimension and number of parameters of the covariance matrix of the random effects for several combinations of sequential times and longitudinal variables

<table>
<thead>
<tr>
<th>Seq. times</th>
<th>Long. variables</th>
<th>$D_u$ Dimension</th>
<th>$D_u$ Parameters</th>
<th># $D_u$ Dimension</th>
<th># $D_{lm}$ Parameter</th>
<th>$\Gamma$ Dimension</th>
<th>Total Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>$2 \times 2$</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>$4 \times 4$</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>$2 \times 2$</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>$6 \times 6$</td>
<td>21</td>
</tr>
<tr>
<td>1</td>
<td>$L$</td>
<td>$2 \times 2$</td>
<td>3</td>
<td>$L$</td>
<td>2</td>
<td>$2L \times 2L$</td>
<td>$3L + 2(L - 1)L$</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$3 \times 3$</td>
<td>6</td>
<td>2</td>
<td>$3 \times 3$</td>
<td>$6 \times 6$</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>$3 \times 3$</td>
<td>6</td>
<td>3</td>
<td>$3 \times 3$</td>
<td>$9 \times 9$</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>$L$</td>
<td>$3 \times 3$</td>
<td>6</td>
<td>$L$</td>
<td>$3 \times 3$</td>
<td>$3L \times 3L$</td>
<td>$6L + \frac{9}{2}L(L - 1)$</td>
</tr>
</tbody>
</table>

Seq. times stands for the number of sequential times and Long. variables is the number of longitudinal variables considered. # $D_u$ represents the number of $D_u$ matrices and # $D_{lm}$, the number of $D_{lm}$ matrices.
Chapter 5

Towards JMseq($K, L$) algorithm: packages JM and JMbayes

5.1 Introduction

The R packages JM (Rizopoulos, 2010) and JMbayes (Rizopoulos, 2014a) for the frequentist and Bayesian approaches of joint modeling have facilitated a lot the application of joint model methodology to real case examples. In this Chapter we aim to study the differences and similarities between both packages in terms of how they work and what we can do with each of them. First we will present the internal structure of their main estimation functions and then detect which parts need to be modified in order to allow the incorporation of more than one time to event and the information on more than one longitudinal variable.

In order to ease reading, in this Chapter we have avoided the technical details of the implementation of the packages, such as the utilization and manipulation of different types of objects, the construction of mechanical procedures to deal with integrals and other complex operations or the implementation of several internal functions that are used in particular moments. Therefore, although we have taken an in-depth look into the implementation, we present the information here without explicitly referring to the code.

In this Chapter we examine the two main functions in JM and JMbayes: jointModel() and jointModelBayes(). Both functions fit shared parameter models for the joint modeling of longitudinal responses and time-to-event data although jointModel() does it under a maximum likelihood approach and jointModelBayes() under a Bayesian perspective.

With all the knowledge acquired from studying the structure of these estimation functions, we will be able to design a procedure towards the estimation of JMseq($K, L$), the joint model function for $K$ sequential survival times and $L$ longitudinal variables.
5.2 Package JM

Introduction

The R package JM has been developed to fit a variety of joint models for normal longitudinal responses and time-to-event data under maximum likelihood. Although the main function in JM package, jointModel(), can account for several issues such as competing risks, different types of censoring or different approximations for the baseline hazard function, we will focus on the structure of the function when fitting a model using splines to approximate the baseline hazard function and assuming that the risk for an event at time $t$ is associated with the subject-specific mean of the longitudinal outcome at the same time point. We study models with these characteristics in order to easily compare functions jointModel() and jointModelBayes().

5.2.1 Base structure

The two main arguments needed to call the model-fitting function are a linear mixed effects object fit as returned by function lme() of package nlme (Pinheiro et al., 2012) and a survival object fit as returned by function coxph() of package survival (Therneau, 2013).

The jointModel() function follows the next algorithm:

1. With the information from the lme and the survival objects, extraction of information and construction of the longitudinal variable $y = X\beta + Zb$.

\[\downarrow\]

2. Construction of a list of information about elements that deal with the control of the function. This list includes information on the iterations for the EM and the Newton algorithm, the tolerance to declare EM convergence, the knots and the order of the spline to approximate the baseline risk function and the number of Gauss-Hermite quadrature points used to approximate the integral with respect to random effects (3.9) and the Gauss-Kronrod points to approximate the integral in the definition of the survival function (3.10).

\[\downarrow\]

3. With the number of points specified in the control argument, computation of the Gauss-Kronrod weights and knots.

\[\downarrow\]
Chapter 5. Towards JMseq($K, L$) algorithm: packages JM and JMbayes

4. Construction of the knots in order to evaluate the design matrix for the B-splines to approximate the log baseline hazard risk function.

5. With the information on the knots, the order of the spline and the times where we want to evaluate it, construction of the design matrix for the B-splines.

6. From the lme model, extraction of the covariance matrix for the random effects.

7. Through initial.surv function (that it will be explained in detail in the next section), construction of the initial values for the survival parameters.

8. Assemblage of the initial values for parameters from both the survival and the longitudinal models.

9. Through function spline.PHGHfit (in this case, because we are approximating the log baseline hazard with splines) and with all the information about the parameters, the initial values and the control arguments that we have obtained in the previous steps of the algorithm, estimation of the parameters of the joint model and their confidence intervals.

10. Construction of the object that the function will return, that will be an object of class jointModel.

5.2.2 Most important procedures

In this section we will discuss in more detail some parts of the jointModel() code. We begin by analyzing its two main internal functions.
initial.surv()

Using information about the different times to event from each individual, the $\beta$ parameters from the estimation of the \texttt{lme} object and the design matrix from the B-splines, this function returns initial estimations (that will be used in the function that computes the final estimations) for parameters $\gamma_h$ in (3.4):

$$
\log h_0(t) = \gamma_{h,0} + \sum_{q=1}^{Q} \gamma_{h,q} B_q(t, \nu)
$$

and parameters $\alpha$ and $\gamma$ in (3.5):

$$
h_i(t | M_i(t), w_i) = \exp\{\gamma^T w_i + \alpha m_i(t)\}.
$$

The function is structured as follows:

1. With the information about the covariates and the survival events, construction of a \texttt{data.frame} with all the information needed to fit a Cox model.
2. Through the function \texttt{paste()}, construction of the formula of the Cox model to be fitted.
3. Fitting of a Cox proportional hazards model with the covariates in the previous \texttt{data.frame}.
4. From the Cox model adjusted, collection of the coefficients for $\alpha$ and $\gamma$.
5. Fit a parametric survival regression model with an assumed Weibull distribution in order to obtain preliminary estimations for the coefficients $\gamma_{h_0}$.
6. Assembling the estimation of the parameters in the object to be returned.

**Estimation Function**

Maximum likelihood estimates are obtained by maximizing the log-likelihood function

$$
\log (L(\theta)) = \sum_i \log p(T_i, \delta_i, y_i, \theta).
$$

The integral in (3.9) is approximated using the Gauss-Hermite rule and the integral in (3.10) is approximated using the Gauss-Kronrod rule (Press et al., 2007). The maximization of the log-likelihood starts with a fixed number of EM iterations. EM algorithm (Dempster et al., 1977) is an iterative method for finding maximum likelihood estimates of parameters. It alternates between an expectation step (E), that creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization step (M), that maximizes the log-likelihood found in the previous step and computes estimates for parameters. These estimates are used to determine the distribution of variables in the next E step. If, after all the EM iterations, convergence is not achieved, the algorithm switches to quasi-Newton iterations until convergence. For the EM iterations, convergence is declared whenever one of the following criteria is satisfied:

$$
\max \left\{ \left| \theta^{it} - \theta^{it-1} \right| / (\left| \theta^{it-1} \right| + \text{tol}_1) \right\} < \text{tol}_2, \\
L(\theta^{it}) - L(\theta^{it-1}) < \text{tol}_3 \left\{ \left| L(\theta^{it-1}) \right| + \text{tol}_3 \right\},
$$

where $\theta^{it}$ denotes the parameter estimates at iteration $it$, $\text{tol}_1$, $\text{tol}_2$, and $\text{tol}_3$ are predefined tolerance levels.
where \( \theta^i \) denotes the parameter values at the \( i \)-th iteration, \( L(\cdot) \) is the log-likelihood function and \( \text{tol}_1, \text{tol}_2 \) and \( \text{tol}_3 \) are specified via the \text{control} argument (default values are \( 10^{-3} \), \( 10^{-4} \) and \( 10^{-8} \), respectively). In the \text{control} argument the user can also specify the optimization routine for the quasi-Newton iterations and the number of quadrature knots for the Gauss-Hermite and the Gauss-Kronrod rules.

### 5.3 Package JMbayes

#### Introduction

The \texttt{R} package \texttt{JMbayes} fits joint models under a Bayesian approach. \texttt{JMbayes} can fit a wide range of joint models, including joint models for continuous and categorical longitudinal responses. Moreover, it provides several options for modeling the association structure between the two outcomes. In many regards, the design of package \texttt{JMbayes} is similar to the one of package \texttt{JM}. Thus, although function \texttt{jointModelBayes()} allows a lot of choices, for example, the type of association between both processes or the estimation of the baseline risk function, to study its structure we are going to focus, as in function \texttt{jointModel()}, on the case in which the baseline risk function is approximated via regression splines and it is assumed a relative risk model as in (3.2), i.e. the risk for an event at time \( t \) is associated with the subject-specific mean of the longitudinal outcome at the same time point.

#### 5.3.1 Base structure

In a similar way as in \texttt{jointModel()} function, the main arguments needed to call the \texttt{jointModelBayes()} function are a linear mixed effects object as the one that fits function \texttt{lme()} of package \texttt{nlme} and a survival object fit by function \texttt{coxph()} or function \texttt{survreg()} of package \texttt{survival}.

With these two main arguments, the function \texttt{jointModelBayes()} is structured in the following steps:

1. With the information from the \texttt{lme} and the survival objects, extraction of information and construction of the longitudinal variable \( y = X \beta + Zb \).

\[ \downarrow \]

2. Construction of a function to define the density of the longitudinal outcome.

\[ \downarrow \]

3. Construction of a function to define the density of the random effects.

\[ \downarrow \]
4. Extraction of the covariance matrix from the random effects model.

5. Construction of the control parameter, which includes information about the iterations in the Markov chain Monte Carlo method (such as the number of total iterations or how many iterations are used for adaptation or are discarded as burn-in) and about the knots and the order of the spline to approximate the log baseline risk function.

6. Calculate the Gauss-Kronrod weights and knots used to approximate the survival function in (3.14).

7. Construct the numeric vector of knots for the spline approximation of the log baseline risk function and evaluate the design matrix for the B-splines. The location of the knots can be based on percentiles of the observed event times $T_i$ or on percentiles of the true event times $\{T_i: T_i^* \leq C_i\}$.

8. Through initSurvival() function (that we will explain in detail in the next section), obtention of the initial values for the survival parameters and their covariances.

9. If the user has not specified user-defined priors, construction of the prior mean vectors and prior precision matrices for $\alpha$, $\gamma$, $\gamma_{h0}$ and for the fixed and random effects. In particular, for the vector of fixed effects of the longitudinal submodel $\beta$, for the regression parameters of the survival model $\gamma$, for the vector of spline coefficients for the baseline hazard $\gamma_{h0}$ and for the association parameter $\alpha$, independent univariate non-informative normal priors are used. For the covariance matrix of the random effects an inverse Wishart prior is assumed.

10. Construction of an object that gathers the information about all the parameters’ covariances.
11. Through function \texttt{MCMCfit()} and with all the information obtained and constructed until this point, estimation of the parameters of the model.

\[ \downarrow \]

12. Construction of the object that the function will return, that will be an object of class \texttt{JMbayes}.

### 5.3.2 Most important procedures

In this section we will discuss in more detail some parts of the \texttt{jointModelBayes()} function. We begin with its two main internal functions.

\texttt{initSurvival()}

This function returns the initial estimation of parameters $\alpha$ and $\gamma$ in (3.5) and $\gamma_{h0}$ in (3.4) and their covariances. The function is structured as follows:

1. With the information received about the covariates and the survival events, construction of a \texttt{data.frame} with all the information needed to fit a Cox model.
2. Fitting of a time-dependent Cox model with the covariates in the previous \texttt{data.frame}.
3. From the coefficients from the Cox model, compilation of the coefficients for $\alpha$ and $\gamma$ and the first estimation for its covariances.
4. Construction of $\gamma_{h0}$ and its covariance matrix via searching the optimal values in the optimization of a function for the B-splines.
5. With matrices multiplications and functions to calculate the integrals in (3.13) and (3.14), definition of a function that computes (3.13). Taken the estimates for $\gamma$, $\alpha$ and $\gamma_{h0}$ as initial parameters, optimization of the function in order to obtain the values that minimize it and use them as the best estimates for the survival parameters and its covariances.
6. Gathering the estimate parameters and their covariances in the object to be returned.

\texttt{MCMCfit()}

This internal function is the implementation of the MCMC algorithm that samples from the posterior conditional distributions of the parameters and the random effects. For the majority of the posterior conditional random walk Metropolis is used with exceptions such as that when the random effects are assumed normally distributed (this can be controlled with argument \texttt{df.RE}), the posterior conditional distribution for the random effects precision matrix $\mathbf{D}^{-1}$ is a Wishart distribution. The implementation uses the information of the separately fitted mixed effects model and Cox models to obtain the covariances matrices of the distributions for the random walk Metropolis algorithm (which is a Markov chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution). On the one hand, for $\mathbf{B}$ and $\mathbf{b}$, the covariance matrices are
taken from the mixed model. On the other hand, for the regression coefficients in the linear predictor of the survival submodel and the B-spline coefficients $\gamma_{ho}$, a two-stage approach is applied.

First of all, during and adaptive phase (default value is 3000), these proposal distributions are tuned and every a pre-specified number of iterations (default is 100) the acceptance rate of the algorithms are checked. Then there is a burn-in period (default is 3000) and after that, the algorithm continues to run for extra iterations (default is 200000). The chains of the parameters are thinned according to an argument specified by the user (default is to keep 2000 iterations for each parameter). Thus, after each calculation, if the iteration has to be kept, the results are stored in its correspondent variables. Although at the end of the process the function returns the means of these variables as well as the variables values, when a `summary()` is executed, only the means are reported.

### 5.4 Similarities and differences between JM and JMbayes

After studying in detail the structure of the estimation functions in JM and JMbayes, in this section we will comment differences and similarities between them. First of all, we will go through the differences in two arguments that we have already brought out (the choice of the approximation of the baseline risk function and the parameterization of the linear predictor for the survival model). Then, comparing the structures of `jointModel()` and `jointModelBayes()`, we will note their main differences.

#### 5.4.1 Baseline risk function

One of the differences between the two packages is the options available for the baseline risk function $h_0(t)$. As we have noted before, although in a Cox Model the baseline risk function can be left unspecified, it is not the case for the joint models so, in order to avoid underestimation of the standard error of the parameter estimates it is necessary to choose a proper $h_0(t)$.

The choice of $h_0$ is specified through the argument `baseHaz` in JMbayes and `method` in JM. In JMbayes we have only two choices, regression splines or penalized splines. For regression splines, the logarithm of the baseline hazard function is expressed as

$$
\log h_0(t) = \gamma_{ho,0} + \sum_{q=1}^{Q} \gamma_{ho,q} B_q(t, \nu)
$$

where $B_q(t, \nu)$ denotes the $q$-th basis function of a B-spline with knots $\nu = (\nu_1, \ldots, \nu_Q)$ and $\gamma_{ho}$ is the vector of spline coefficients. Note that as the number of knots increases, the specification of the baseline hazard becomes more flexible. The penalized version can be fitted by specifying for $\gamma_{ho}$ the improper prior (Rizopoulos, 2014b)

$$
p(\gamma_{ho} | \tau_h) \propto \exp \left( -\frac{\tau_h}{2} \gamma_{ho}^T K \gamma_{ho} \right)
$$

(5.1)
where \( \tau_h \) is the smoothing parameter, \( K = \Delta^T \Delta_r \) where \( \Delta_r \) denotes the differences matrix and \( \rho(K) \) is the rank of \( K \).

Whereas the argument `baseHaz` in `jointModelBayes()` specifies only the type of survival model to be fitted, the argument `method` in the function `jointModel()` gives information not only about the survival model but also about the numerical integration method to approximate the integral in (3.9). Available options are:

- "weibull-AFT-GH" that assumes the Weibull model under the accelerated failure time or the relative risk formulation.
- "piecewise-PH-GH" which assumes the relative risk model with a piecewise risk function as the one we defined in Chapter 3:

\[ h_0(t) = \sum_{q=1}^{Q} \xi_q I(\nu_{q-1} < t < \nu_q), \]

where \( 0 = \nu_0 < \nu_1 < \cdots < \nu_Q \) is a split of the time scale (with \( \nu_Q \) being larger than the largest observed time) and \( \xi_q \) denotes the value of the hazard in the interval \( (\nu_{q-1}, \nu_q) \).
- "Cox-PH-GH" that uses a relative risk model with an unspecified baseline risk function. This option is the one proposed by Wulfsohn and Tsiatis (1997).
- "spline-PH-GH" that assumes a relative risk model with a spline-approximated baseline function. The baseline risk function \( \log h_0(t) \) is expanded as in (3.4).

The string "GH" indicates the type of Gauss-Hermite rule to be used to approximate (3.9). If the chosen option for the method contains the string "GH", the standard Gauss-Hermite quadrature is used, whereas if the string is "aGH", the type of rule used is the pseudo adaptive Gauss-Hermite (Rizopoulos, 2013). The only option that does not use the Gauss-Hermite rule is "ch-Laplace". This option assumes a similar survival model as the one in "spline-PH-GH" but the approximation of the integral in (3.9) is done by using a fully exponential Laplace transformation. This is useful in settings where the subject-specific longitudinal profiles are nonlinear and thus, modeling them using Gauss-Hermite can be time consuming due to the high dimension of the random effects. The Laplace method gives a quite reasonable approximation in less computing time.

### 5.4.2 Linear predictor for the survival submodel

Both `JM` and `JMbayes` allow the choice of the linear predictor for the survival model (through argument `param` in `JMbayes` and `parameterization` in `JM`). Thus, the linear predictor for the survival model is written as

\[ \gamma^T w_i + f(m_i(t), m'_i(t), b_i; \alpha, \alpha_s). \] (5.2)

In particular, through the specification of function \( f(m_i(t), m'_i(t), b_i; \alpha, \alpha_s) \), the options available for the association are:
Chapter 5. Towards JMseq($K, L$) algorithm: packages JM and JMbayes

\[
\hat{f}(m_i(t), m'_i(t), b_i; \alpha, \alpha_s) = \alpha f_1(m_i(\text{max}(t - k), 0)) \text{ for } \text{param = td-value} \text{ and parameterization = value}.
\]

\[
\hat{f}(m_i(t), m'_i(t), b_i; \alpha, \alpha_s) = \alpha_s f_2(m'_i(\text{max}(t - k), 0)) \text{ for } \text{param = td-extra} \text{ and parameterization = slope}.
\]

\[
\hat{f}(m_i(t), m'_i(t), b_i; \alpha, \alpha_s) = \alpha f_1(m_i(\text{max}(t - k), 0)) + \alpha_s f_2(m'_i(\text{max}(t - k), 0)) \text{ for } \text{param = td-both} \text{ and parameterization = both}.
\]

\[
\hat{f}(m_i(t), m'_i(t), b_i; \alpha, \alpha_s) = \alpha T b_i \text{ for } \text{param = shared-RE} \text{ (option not available in JM)}.
\]

\[
\hat{f}(m_i(t), m'_i(t), b_i; \alpha, \alpha_s) = \alpha T (\beta + b_i) \text{ for } \text{param = shared-betasRE} \text{ (option not available in JM)}.
\]

where in all these cases $k$ is specified by the lag argument, $f_1(\cdot)$ and $f_2(\cdot)$ denote possible transformation functions, $b_i$ denotes the vector of random effects for the $i$-th subject and $\beta$ the fixed effects.

5.4.3 Structure differences

As we have already said, the design of package JMbayes is similar to the one of package JM. Therefore, although they fit joint models under different perspective, there are a lot of common points between them. In this section, we will make a compendium of the differences in the algorithms in their estimating functions, which we have already explained in the previous section. Thus, the main differences between functions jointModel() and jointModelBayes() are:

- The arguments in the control element. The two packages have different ways to estimate parameters; thus, the elements they need to control the estimations’ procedures are different. On the one hand, for the JM case, the iterations for the EM and quasi-Newton algorithms or the tolerance to declare convergence are needed. On the other hand, JMBayes has to keep track of the iterations in the MCMC method. There are also common parameters for both functions, such as the knots and the order of the spline use to approximate the log baseline risk function.

- When computing the initial estimates for the survival parameters, whereas for $\alpha$ and $\gamma$ the method is the same in both functions, to give initial estimates for $\gamma_{h0}$, the functions proceed in different ways: jointModel() fits a survival regression model with a Weibull distribution and jointModelBayes() constructs functions to compute the integrals in the likelihood estimation and optimize them in order to obtain the $\gamma_{h0}$ estimates.

- The main difference between both packages is the procedure used to estimate the parameters. Whereas the frequentist approach uses an EM algorithm, the Bayesian one takes profit of the Monte Carlo Markov Chain method. In this step is where the parameters in the control argument are used. Moreover, for the Bayesian estimation it is also necessary to construct the prior mean vectors and the prior precision matrices for all parameters in the model.
5.5 Extension towards JMseq($K,L$)

To extend the procedures analyzed in this Chapter to the case of $K$ sequential survival times and $L$ longitudinal variables, we are going to focus on the Bayesian perspective because, as long as we know, this option is more flexible than the frequentist one in terms of introducing changes.

One of the most important issues is to detail the input objects for the estimation function, because these objects are the ones that provide the majority of the information needed to fit the joint model. Whereas in the `jointModelBayes()` function the inputs were an `lme` object and a Cox model, in the case we are studying arises the necessity to include much more information through these initial objects:

- To include the information about the $L$ longitudinal variables, we will need to fit $L$ different `lme` models, one for each longitudinal variable. Because these `lme` objects will assume that variables are independent, we have to implement a procedure to account for the correlation between them. In other words, we have to manually construct the matrix that accounts for the correlation between variables (we have explicitly written this $\Gamma_{K,L}$ matrix in Chapter 4). Thus, the joint model function we are implementing should have an argument to specify the structure of the covariance matrix for each couple of variables because, in general, we can not assume that every pair of longitudinal variables have the same correlation. If the user does not specify this structures, as a default value we could assume that the matrices are unstructured, and estimate all the parameters in them (see Table 4.1 for more details on the number of parameters for $\Gamma_{K,L}$) but this will lead to a very computationally demanding procedure. On the other hand, if the user does not specify which relation exists between the variables, we can assume that their random effects are independent and, thus, construct the block matrices outside the diagonal of $\Gamma_{K,L}$ as matrices filled with zeros. With the specification of the relation between all the pairs of variables, we will have the complete information about the longitudinal part of the model that we need in order to estimate the joint model.

- For the survival part, we need to fit $K$ Cox models, one for each sequential time. The Cox models will model gap times $t_{k,i}$, $k = 1, \ldots, K$, with information about baseline covariates (covariates that can be different for different times) and information about the previous times, because their length can have an effect on the estimation of the following times. Thus, for example, the Cox model for $t_{3,i}$ will take information from the baseline covariates that have an effect on the third interval of time as well as information from $t_{1,i}$ and $t_{2,i}$. This information about $t_{1,i}$ and $t_{2,i}$ will be acknowledged with parameters $\delta_{3,1}$ and $\delta_{3,2}$. These $\delta$ parameters can explain both correlation between one time and the previous observed times and also the potential mechanism of censoring that can affect the relation between one time and the preceding ones.

Once we have determined the objects that will be used as main arguments in the function to estimate JMseq($K,L$), we are going to detail its algorithm. The structure will be similar to the one in the package `JMbayes` but will include some modifications to deal with the incorporation of sequential times and several longitudinal variables.
Chapter 5. Towards JMseq(K, L) algorithm: packages JM and JMbayes

The steps of the algorithm will be:

1. In the same way as we did in `jointModelBayes()`, with the information from the \( L \) lme and \( K \) survival objects, extraction of information and construction of the \( L \) longitudinal variables \( y = X\beta + Z\theta \).

2. Construction of functions to determine the density of the longitudinal outcomes. In this case, we will have \( L \) different functions because we cannot assume that all \( L \) variables follow the same distribution.

3. In a similar way, construction of the function to compute the density of the random effects for all longitudinal variables. Again, we will have \( L \) different functions, each one adjusting the density of the random effects of one of the variables.

4. Extraction of the \( L \) covariance matrices from the lme objects. These matrices are the matrices in the diagonal of \( \Gamma_{K,L} \). With the information the user has provided about the covariance between all pairs of longitudinal variables, we can construct the matrices outside the diagonal. Therefore, at this point we have enough information to explicitly construct \( \Gamma_{K,L} \) in order to estimate its parameters.

5. In order to have some control over the function, we have to construct an object to gather information about, for example, the iterations in the Markov chain Monte Carlo method or the knots and the order of the spline to approximate the baseline hazard function.

6. As we did in the function to estimate JMseq(1, 1), computation of Gauss-Kronrod weights and knots used to approximate the survival function.

7. Construction of the numeric vector of knots for the spline approximation of the logarithm of the baseline risk function. In this case, since knots depend on times, we will have a pack of knots for each sequential time. Using these different knots to evaluate the design matrix for the B-splines, we will get \( K \) different design matrices.
8. In this next step, similarly as `initSurvival()` in `jointModelBayes()` and `init.surv()` in `jointModel()`, we have to construct an internal function to obtain initial estimations for the parameters of the survival part. We have to proceed in a similar way as we did in previous steps, extending the procedure for one survival time to \( K \).

First of all, we have to construct \( K \) different data.frames with all the information needed to fit a Cox Model for each sequential times. Because each data.frame contains different information, the Cox Model fitted for each sequential time will provide different estimations. These estimations will be the estimates of parameters \( \gamma_k, \alpha_{k,l} \) and \( \delta_{k,1}, \ldots, \delta_{k,k-1} \) with \( k = 1, \ldots, K \) and \( l = 1, \ldots, L \) from formulas in Section (4.4.2) and their covariances.

To obtain initial estimations for the \( \gamma_{h_0} \) parameters in the approximation of the logarithm of the baseline risk function, we will use the function define in `initSurvival()` and use its optimal values as our \( \gamma_{h_0} \) estimations. We will have to run the function \( K \) times because the estimations will be different for each \( h_0 \).

Finally, analogously as we did in JMseq(1,1), with the initial estimations of the parameters for the Cox model, and adapting the function implemented in `initSurvival()` to include information from the \( L \) longitudinal variables, we will find the optimal values of the function and use them as best estimates for the survival parameters and its covariances.

9. Construction of the prior means and prior precision matrices for \( \alpha_{k,l}, \gamma_k, \gamma_{h_0} \) and for the fixed and random effects of all variables. For the parameters that JMseq(1,1) and JMseq(\( K, L \)) share, we will assume the same distributions as in `jointModelBayes()`.

For the new parameters, \( \delta_{k,1}, \ldots, \delta_{k,k-1} \) for \( k = 1, \ldots, K \) we assume also independent univariate non-informative normal priors whereas for the covariance matrices between the random effects of different longitudinal variables (matrices outside the diagonal of \( \Gamma_{K,L} \)) we would assume distributions that capture the dependency between the \( L \) different lme objects.

10. Gathering of all parameters’ covariances in an object that will be used when calling function `MCMCfit()`.
11. `MCMCfit()` function will estimate the parameters of the model, running as much iterations as we have defined in `control` argument and using the priors defined to get the initial estimations for the parameters.

↓

12. Construction of the object that the function will return, that will be an object of class `JMseq`. 
Chapter 6

Discussion and future research

6.1 Discussion and conclusions

In clinical studies, it is increasingly common to measure the association between a longitudinal variable and the risk for an event. Because these longitudinal variables are usually time-dependent, separate analysis of survival and longitudinal data can lead to biased results. Therefore, a jointly modelization of both approaches is needed in order to correctly acknowledge the link between both survival and longitudinal processes.

Joint modeling can be approached from frequentist and Bayesian perspectives and in this work we have highlighted that, although they have several points in common, they have major differences in terms of estimation. From the joint modeling perspective, most of the work has focused on models with one longitudinal outcome and one survival time. Nevertheless, an issue that often arises in studies on health-related disciplines is that patients generate a series of outcomes that can be used as longitudinal variables. Moreover, it is also of interest to study multiple survival times for each patient, such as sequential survival times.

The main goal of this work has been to understand joint modeling from frequentist and Bayesian points of view and to develop a methodology able to incorporate in a joint model sequential survival times and information from more than one longitudinal covariates. This work was motivated by the TIBET clinical trial, that contemplated the incorporation of interruption periods in the administration of an intensive therapy for HIV infected patients, which generated a sequence of survival times for each subject under study. Because we have information on times without treatment and times with treatment along with the measurements of CD4 cell counts and viral load, this leads us to a joint modeling with $K$ sequential survival times and $L$ longitudinal variables, what we have denoted JMseq ($K, L$).

Firstly, we have presented the joint modeling approach for one survival time and one longitudinal covariate and extended it for what we have called JMseq ($K, L$). After that, we have illustrated how R packages JM and JMbayes work by showing an application to the TIBET study, concluding that both approaches lead to similar results. Then, working towards the implementation of a procedure able to compute estimates for JMseq ($K, L$) and taking advantage of the already developed R packages JM and JMbayes, we have analyzed the estimation functions in both packages and pointed out what parts of the
implementation should be modified and how to modify them to incorporate the effect of $K$ sequential survival times and $L$ longitudinal variables.

After studying and comparing the frequentist and Bayesian approaches, we have reach the conclusion that, in terms of extending the already implemented procedures, the Bayesian one allows more flexibility to include more survival times and more longitudinal variables. Thus, in a future implementation of a procedure for JMseq($K, L$) the focus will be on implementing a procedure similar to jointModelBayes().

To conclude, with the development of the methodology for JMseq($K, L$) the profits of a sequential approach when having sequential survival times and longitudinal variables have been explored. Moreover, we have experimented the meticulous work that is implementing a procedure in terms of programming as well as the computational issues that can arise.

### 6.2 Future research

The research that has been carried out in this master’s degree thesis opens some areas of interest in the joint modeling area that deserve further attention. They can be summarized as follows:

- The work we have developed in this master’s degree thesis can be considered a first step towards the implementation of a function that, in a similar way to the function jointModelBayes(), would fit joint models for sequential survival times and several longitudinal variables. Thus, using what is already implemented, the next step would be the implementation in R of a joint model function for $K$ sequential survival times and $L$ longitudinal variables.

- Although we have assumed that the risk for an event at time $t$ is associated with the subject-specific mean of the longitudinal outcome at the same time point, it would be reasonable to consider that also the slope of the longitudinal trajectory at this time has an effect on the hazard function. Therefore, we could explore different associations between the risk for an event and the longitudinal outcome.

- Another issue is that, once the joint models have been validated, we can work with the powerful tool already available in JM and JMbayes packages that derives results in terms of survival predictions. Thus, with the set of longitudinal measurements and the vector of baseline covariates for a new subject, the goal would be to predict conditional probability of surviving time for that particular subject. This approach is towards personalized medicine, an issue that has capture professionals attention in recent years.
Bibliography


Appendix A

R code

```r
> ##############
> ##LIBRARIES##
> ##############
> library(survival)
> library(nlme)
> library(JM)
> library(JMbayes)
> library(xtable)
>
> # FUNCTIONS TO COMPUTE CONFIDENCE INTERVALS#
> confint2log <- function(jm, alpha=0.05){
> + ## sigma
> + se.logsigma <- sqrt(vcov(jm)["Y.sigma", "Y.sigma"])
> + sigmahat <- jm$coefficients$\sigma
> + se.sigma <- se.logsigma\*sigmahat
> + ic.sigma <- c(sigmahat-qnorm(1-alpha/2)*se.sigma,
> +                   sigmahat+qnorm(1-alpha/2)*se.sigma)
> + return(ic.sigma)
> + }
> confint2quadrat <- function(sigmabhat, se.sigmab2, alpha=0.05){
> + ## se(\sigma b)
> + se.sigmab <- se.sigmab2/(2*sigmabhat)
> + ic.sigmab <- c(sigmabhat-qnorm(1-alpha/2)*se.sigmab,
> +                   sigmabhat+qnorm(1-alpha/2)*se.sigmab)
> + return(ic.sigmab)
> + }
```
## MODELS FOR T1 AND CD4 ##

### read data ###
```r
dades <- read.table("dades_T1CD4.txt", sep="\t", head=T)
dades.id <- dades[!duplicated(dades$ID),]
```

### survival model ###
```r
coxfit <- coxph(Surv(T1,d)~log10(VLpret), data=dades.id, x=TRUE)
```

### lme model ###
```r
lmefit <- lme(sqrt(CD4)~obstime, random=~obstime|ID, data=dades)
```

### joint model ###
```r
jm <- jointModel(lmefit, coxfit, timeVar="obstime",
+ method="spline-PH-aGH")
```

### joint model bayes ###
```r
jmb <- jointModelBayes(lmefit, coxfit, timeVar="obstime",
+ baseHaz="regression-splines")
```

### confidence intervals ###
```r
confint(jm)
confint(jmb)
xtable(jm)
xtable(jmb)
confint2log(jm)
confint2log(jmb)
```
> # MODELS FOR T1 AND VIRAL LOAD #
> #read data
> dades <- read.table("dades_T1CV.txt", sep="\t", head=T)
> dades.id <- dades[!duplicated(dades$ID),]
> #survival model
> coxfit <- coxph(Surv(T1,d)~sqrt(CD4N),data=dades.id,x=TRUE)
> #lme model
> lmefit <- lme(log10(CV)~obstime,random=~1|ID, data=dades)
> #joint model
> jm <- jointModel(lmefit,coxfit,timeVar="obstime",
+                method="spline-PH-aGH")
> summary(jm)
> #joint model bayes
> jmb <- jointModelBayes(lmefit,coxfit,timeVar="obstime",
+                        baseHaz = "regression-splines")
> summary(jmb)
> #confidence intervals
> confint(jm)
> confint(jmb)
> xtable(jm)
> xtable(jmb)
> confint2log(jm)
> confint2log(jmb)
> ############################################################################
> ## MODELS FOR T1+T2 AND CD4 ##
> ############################################################################
> ## read data
> dades <- read.table("dades_T1T2CD4.txt", sep="\t", head=T)
> dades.id <- dades[!duplicated(dades$ID),]
> ## survival model
> coxfit <- coxph(Surv(T1T2,d)~log10(VLpret), data=dades.id, x=TRUE)
> ## lme model
> lmefit <- lme(sqrt(CD4)~obstime, random="obstime|ID", data=dades)
> ## joint model
> jm <- jointModel(lmefit, coxfit, timeVar="obstime",
> + method="spline-PH-aGH")
> summary(jm)
> ## joint model bayes
> jmb <- jointModelBayes(lmefit, coxfit, timeVar="obstime",
> + baseHaz = "regression-splines")
> summary(jmb)
> ## confidence intervals
> confint(jm)
> confint(jmb)
> xtable(jm)
> xtable(jmb)
> confint2log(jm)
> confint2log(jmb)
> ###########################################################################
> ## MODELS FOR T1+T2 AND VIRAL LOAD ##
> ###########################################################################
> ##read data
> dades <- read.table("dades_T1T2CV.txt", sep="\t", head=T)
> dades.id <- dades[!duplicated(dades$ID),]
> ##survival model
> coxfit <- coxph(Surv(T1T2,d)~sqrt(CD4N), data=dades.id, x=TRUE)
> ##lme model
> lmefit <- lme(log10(CV)~obstime, random=~1|ID, data=dades)
> ##joint model
> jm <- jointModel(lmefit, coxfit, timeVar="obstime", 
> + method="spline-PH-aGH")
> summary(jm)
> ##joint model bayes
> jmb <- jointModelBayes(lmefit, coxfit, timeVar="obstime", 
> + baseHaz = "regression-splines")
> summary(jmb)
> ##confidence intervals
> confint(jm)
> confint(jmb)
> xtable(jm)
> xtable(jmb)
> confint2log(jm)
> confint2log(jmb)