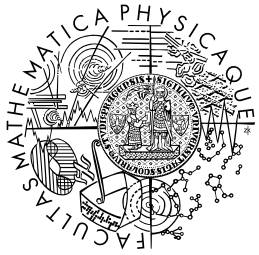

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Jana Vorlíčková

Joint Models for Longitudinal and Time-to-Event Data

Department of Probability and Mathematical Statistics

Supervisor of the master thesis:	doc. RNDr. Arnošt Komárek, Ph.D.
Study programme:	Mathematics
Study branch:	Probability, Mathematical Statistics and Econometrics

Prague 2020

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ..... signature of the author

I am especially grateful to the supervisor doc. RNDr. Arnošt Komárek, Ph.D. for his time, constructive remarks and helpful comments that improved the quality of the thesis. I would also like to thank my family and friends for their support and patience during my studies.

Title: Joint Models for Longitudinal and Time-to-Event Data

Author: Jana Vorlíčková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The joint model of longitudinal data and time-to-event data creates a framework to analyze longitudinal and survival outcomes simultaneously. A commonly used approach is an interconnection of the linear mixed effects model and the Cox model through a latent variable. Two special examples of this model are presented, namely, a joint model with shared random effects and a joint latent class model. In the thesis we focus on the joint latent class model. This model assumes an existence of latent classes in the population that we are not able to observe. Consequently, it is assumed that the longitudinal part and the survival part of the model are independent within one class. The main intention of this work is to transfer the model to the Bayesian framework and to discuss an estimation procedure of parameters using a Bayesian statistic. It consists of a definition of the model in the Bayesian framework, a discussion of prior distributions and the derivation of the full conditional distributions for all parameters of the model. The model's ability to estimate the composition of the population with respect to latent classes and estimate the parameters of the model is evaluated in a simulation study.

Keywords: Bayesian statistics, joint model, Cox model, linear mixed effects model, latent class model

Název práce: Sdružené modely pro longitudinální a cenzorovaná data

Autor: Jana Vorlíčková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Metody zabývající se sdruženými modely pro longitudinální a cenzorovaná data umožňují analyzovat tyto dva typy dat souběžně v rámci jednoho modelu. V této oblasti se často využívá propojení lineárního modelu se smíšenými efekty a Coxova modelu skrze latentní proměnnou. V práci jsou prezentovány dva speciální případy, sdružený model se sdílenými náhodnými efekty a sdružený model s latentními třídami. Hlavní pozornost je věnována sdruženému modelu s latentními třídami. Tento model předpokládá existenci skrytých skupin v populaci, které jsou do modelu zaneseny pomocí diskrétní latentní proměnné. Následně předpokládáme, že část modelu analyzující longitudinální data je nezávislá na analýze cenzorovaných dat v rámci jedné třídy. Cílem této práce je představit model v kontextu Bayesovské statistiky a zaměřit se na odhadování parametrů modelu pomocí Bayesovských metod. Diskutujeme volby apriorních rozdělení a poskytujeme odvození plně podmíněných rozdělení pro všechny parametry modelu. Schopnost odhadnutí rozložení skrytých tříd v rámci populace a odhad parametrů modelu je otestována v simulační studii.

Klíčová slova: Bayesovská statistika, sdružený model, Coxův model, lineární model se smíšenými efekty, model s latetními třídami

Contents

Notation **2**

Introduction **3**

1 Preliminaries **5**
1.1 Analysis of longitudinal data 5
1.1.1 Linear model with mixed effects 5
1.2 Analysis of time-to-event data 7
1.2.1 Time-to-event data 7
1.2.2 The Cox model 9
1.3 Bayesian statistics 11
1.3.1 Basic concepts 11
1.3.2 Estimation methods 12

2 Joint Models for Longitudinal and Time-to-Event Data **15**
2.1 Joint models with shared random effects 15
2.1.1 A definition of the model 16
2.2 Joint latent class models 17
2.2.1 Definition of the model 17

3 Estimation of parameters in JLCM **20**
3.1 The maximum likelihood method 20
3.2 The Bayesian approach 21
3.2.1 Prior distributions 23
3.2.2 Posterior distribution 26
3.2.3 A selection of number of classes 27

4 Derivation of full conditional distributions **28**
4.1 The longitudinal model 28
4.2 The survival model 34
4.3 Class probability model 37
4.3.1 Modeling class probabilities 38

5 A simulation study **42**
5.1 A description of the situation 42
5.2 Results 44
5.3 Discussion 46

Conclusion **48**

Bibliography **49**

A Attachments **52**
A.1 Probability distributions 52
A.2 Matrix Algebra 52

Notation

$\boldsymbol{a}, \boldsymbol{A}$	vector
$\mathbb{A}$	matrix
$\lvert\mathbb{A}\rvert$	determinant of matrix $\mathbb{A}$
$\mathbb{A}^\mathsf{T}$	transpose of a matrix $\mathbb{A}$
$\boldsymbol{a}^\mathsf{T}$	transpose of a vector $\boldsymbol{a}$
$\mathbb{A}^{-1}$	inverse of a matrix $\mathbb{A}$
$\mathrm{tr}(\mathbb{A})$	trace of a matrix $\mathbb{A}$
$\mathrm{diag}(a_1, \ldots, a_n)$	diagonal matrix with elements $a_1, \ldots, a_n$ on the diagonal
$\mathbb{I}_n$	$(n \times n)$-identity matrix
$\mathbb{1}(A = a)$	indicator function, is equal to 1 if $A = a$, zero otherwise
$p(\boldsymbol{y})$	density of $\boldsymbol{y}$
$p(\boldsymbol{y}\lvert\boldsymbol{x})$	density of $\boldsymbol{y}$ given $\boldsymbol{x}$

Introduction

The longitudinal data and the survival data are two types of data that are commonly treated in statistical analysis. We typically meet them in medical studies or in relation to biological disciplines. Over the years, several methods have been developed to work with this type of data. To build a model and estimate parameters in an efficient way, and a statistical theory has been created behind these methods. The standard tool to analyze longitudinal data is a linear mixed effect model, a generalized linear mixed model or a generalized estimation equations method. In survival analysis we often meet the Cox model, or the additive hazard regression model etc.

Recently, new types of models were introduced, such as the joint models for longitudinal data and time-to-event data (Faucett and Thomas [1996], Wulfsohn and Tsiatis [1997]). By joint modelling we understand an interconnection of the separate models for both types of data into one complex model. It is another approach how to evaluate a relationship between time-dependent covariates and risk of an event. It also allows us to build one model for the population with latent classes and differentiate between the effects of covariates to the risk of the event and the longitudinal marker for these classes that are not observed.

In this thesis we focus on the models where the relationship between the longitudinal marker $\boldsymbol{Y}$ and the event time T is captured by a latent variable $\boldsymbol{U}$. Special examples of this model are called joint models with shared random effects (SREM), where the latent variable corresponds to random effects, and joint latent class models (JLCM), where the latent variable captures information about class membership of the individual subjects.

These models are introduced in the thesis and then we discuss JLCM in more detail. Commonly, the frequentist approach is used and the parameters of the model are estimated by the maximum likelihood method. However, in this thesis we target to the estimation of the parameters using the Bayesian method including a derivation of full conditional distributions for parameters of the model.

The thesis is structured as follows: In Chapter 1 we go over the basic concept of the theory that is needed to build a theory for joint models. The analysis of the longitudinal data using a linear mixed effect model and the analysis of time-to-event data using the Cox model is shortly introduced. Also, we provide the reader with a brief overview of Bayesian statistics introducing basic terms and a few notes about estimation methods in Bayesian statistics.

Next, in Chapter 2, we move on to the definition of joint models for longitudinal and time-to-event data where we acquaint the reader with the two most common approaches, namely, joint models with shared random effects and joint latent class model. In Chapter 3, the second mentioned model is introduced in more detail and we describe methods on how to estimate the parameters in such a model. This thesis focuses on the Bayesian approach. Thus, the model is properly specified in a Bayesian way of thinking and the choices of prior distributions are discussed.

As the main intention of this thesis is to explore the Bayesian estimation method for joint models detailed derivations of full conditional distributions are presented in Chapter 4. Finally, a short simulation study is presented and we evaluate the performance of Bayesian estimation methods for latent class joint models in Chapter 5.

1. Preliminaries

The method of joint models allows us to build a model using two different types of data, for example the longitudinal data and time-to-event data, combining methods that are typically used separately. To introduce this modeling framework, we first briefly describe the two building blocks of the joint models. These models are based on a link of regression models with mixed effects and survival models.

Moreover, the basics of Bayesian statistics are introduced at the end of this chapter as this is a main method considered in this thesis while we estimate the parameters of interest.

1.1 Analysis of longitudinal data

A situation where the classical linear regression approach is not satisfactory is generated when the collected data is correlated. We concentrate on one type of correlated data, longitudinal data, to be able to model the evolution of factors influencing the survival outcome over time. This data is a result of repeatedly measured characteristics of one subject over time (e.g. subject is a patient in a clinical study). Due to the correlation structure the independence of observations cannot be assumed and the simple regression approach is not appropriate for this type of data. In the following section, a linear mixed-effects model is presented, which is the most common type of model for such a situation.

1.1.1 The linear model with mixed effects

Linear models with mixed effects are based on the linear regression method. Nevertheless, the collected data is no longer assumed to be independent. To solve this complication, we need to add random effects to a set of fixed effects that somehow capture dependency in our observations. These random effects are unobserved. A widely used assumption is that the random effects follow a multivariate normal distribution with a zero mean and a covariance matrix $\mathbb{D}$ that describes the correlation structure within observations measured on one subject.

We define the model properly. We assume that a response follows a continuous distribution. Let us have K independent subject, $i = 1, \ldots, K$ and denote $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^{\mathsf{T}}$ as a response vector for i-th subject and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^{\mathsf{T}}$ is a vector of errors for i-th subject. The assumption of independence of the subjects holds throughout the thesis. Let $\mathbb{X}_i$ be an $n_i \times p$ design matrix for fixed effects, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^{\mathsf{T}}$ is a p-vector of fixed effects, let $\mathbb{Z}_i$ be an $n_i \times q$ design matrix for random effects.

Definition. The responses $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$ satisfy a (single-level) *linear mixed effects model* (LME model) if

$$\boldsymbol{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, K, \tag{1.1}$$

where $\boldsymbol{b}_i$ (the random effects) are independent identically distributed vectors,

$$\boldsymbol{b}_i \sim \mathsf{N}_q(\boldsymbol{0}, \mathbb{D}),$$

$\boldsymbol{\varepsilon}_i$ are independent vectors,

$$\boldsymbol{\varepsilon}_i \sim \mathsf{N}_{n_i}(\mathbf{0}, \sigma^2 \mathbb{I}_{n_i}),$$

and $(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_K)^\mathsf{T}$ are independent of $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K)^\mathsf{T}$.

Consequently, it comes from the definition and equation (1.1) that the linear mixed effect model (Laird and Ware [1982]) can be written as

$$\boldsymbol{Y} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{1.2}$$

where $n = \sum_{i=1}^{K} n_i$, $\boldsymbol{Y}$ is a response vector for all subjects, $\mathbb{X}$ is an $n \times p$ regression matrix, and $\boldsymbol{\Sigma}$ is a block-diagonal matrix with blocks $\Sigma_1, \ldots, \Sigma_K$, such that $\Sigma_i = \mathbb{Z}_i \mathbb{D} \mathbb{Z}_i^\mathsf{T} + \sigma^2 \mathbb{I}_{n_i}$. The implication in the opposite direction from (1.2) to (1.1) does not hold.

While the fixed effects describe an average population effect on the other side the random effects express the subject specific effect within the population, in other words they describe an individual trajectory for each specific subject over time. Therefore, compared to the linear regression model, we need to estimate not only fixed effects $\boldsymbol{\beta}$ but also components of the covariance matrix $\mathbb{D}$, σ^2, and the latent variable $\boldsymbol{b}$. In some cases, however that is not necessary. Note that for the latent variable $\boldsymbol{b}$ we talk about a prediction rather then about an estimation. The estimate of $\boldsymbol{\beta}$ can be obtained using a maximum likelihood method or by derivation of estimates from the so-called Henderson‘s mixed model equations (Henderson [1984]). Assuming that all the inverses exist and $\boldsymbol{\Sigma}$ is known, i.e. σ^2 and all components of covariance matrix $\mathbb{D}$ are known, there exists a closed-form solution for $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \mathbb{X})^{-1} (\mathbb{X}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}),$$

and under same assumptions, the best linear unbiased prediction (BLUE) of $\hat{\boldsymbol{b}}$ takes form,

$$\hat{\boldsymbol{b}} = \mathbb{D}_* \mathbb{Z}^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}),$$

where $\mathbb{D}_*$ is a block-diagonal matrix with K blocks $\mathbb{D}$. We ofteh also need to estimate the components of the covariance matrix $\mathbb{D}$ and σ^2 are needed to be estimated as they are not known. Then the estimates are plugged into the estimator for $\boldsymbol{\beta}$ and $\boldsymbol{b}$. The estimates of the variance parameters could be derived by the restricted maximum likelihood method (REML) (Patterson and Thompson [1971]) or by the maximum likelihood method. In this thesis, the advantages or disadvantages of these approaches will not be discussed, as we are going to focus on Bayesian methods and there is a lot of literature discussing these methods in more detail (e.g. Jennrich and Schluchter [1986], Lindstrom and Bates [1988], and Harville [1974]).

The simple single-level linear mixed effects model is able to be extend to a multi-level linear model with mixed effects. The more complex models allow cross effects and nested random effects. Although the use of extension to the multi-level models is not very common in joint modeling.

1.2 Analysis of time-to-event data

In this section we provide a summary of the basic concepts of analysis of time-to-event data and proportional risk models that play an important role in joint models for longitudinal and survival data. Moreover, we explain the difference in endogenous and exogenous time-varying covariates and provide the motivation for the development of joint models from this point of view.

1.2.1 Time-to-event data

Let $T^* \geq 0$ be the true failure time and it is the variable of interest. In case that we are able to observe the event for each subject, we get a random sample of $T_1^*, \ldots, T_n$. In fact it is quite rare and we are not always able to observe the exact time T^* directly. There are several reasons why this problem arises, such as the length of the experiment, financial costs, or simply a specific type of illness (e.g. reoccurrence of cancer) that may not occur until the end of the patient's life, thus the event is never observed and death is a so-called censoring time.

Then $C \geq 0$ expresses the length of the observation of the subject and it is called censoring time. Depending on the mutual positioning on the time axis of censoring time and the true time to event, the censoring is classified as right, left or interval censoring. The second categorization of censoring is in regard to whether the probability of censoring depends on the failure process. In other words, we distinguish between informative censoring that corresponds to the missing not at random (MNAR) missing data mechanism in longitudinal studies and noniformative censoring, i.e. a dropout of the subject from the study is not related to the expected failure time. This type of censoring is similar to that of the missing completely at random (MCAR) data mechanism.

When it comes to the distribution of random variables $C_1, \ldots, C_n$, it is generally allowed that every variable C_i may follow a different distribution. It is called the random censorship model. Two special cases of this model are mentioned below.

Type I censoring The time $\tau > 0$ of censoring is chosen in advance, i.e. all censoring variables are equal to the same constant, $C_i = \tau$ for all i almost surely. τ indicates the duration of the observation.

Type II censoring The number of failures is chosen in advance. When this number is reached all remaining observations are censored, i.e., $C_i = T^*_{(k)}$ for all i and T^*_k is the k-th order statistic of the random sample $T_1^*, \ldots, T_n^*$.

However, in many real life applications, e.g. clinical studies, both of these types of censoring are unrealistic.

Suppose we have a sample $(T_1^*, C_1), \ldots, (T_K^*, C_K)$, i.e. there is a pair of variables, but only one of the variables is observed for each subject. The two situations arise: $T_i^* \leq C_i$ then the event was observed, on the other hand if $T_i^* > C_i$, the only thing which is known is that the event did not occur during the period of observing and just the censoring time is known.

Instead of (T_i^*, C_i) another type of notation is often used. Let $\delta = \mathbb{1}(T \leq C)$ be the event indicator and $T = \min(C, T^*) = T^* \wedge C$ denotes the time of either

an occurrence of the event or censoring. Thus, our sample contains pairs of observations $(T_1, \delta_1), \ldots, (T_K, \delta_K)$.

Now we define several functions that are commonly employed in the analysis of time-to-event data. First, we introduce the survival function as a probability of surviving time t.

Definition. The function $S(t) = 1 - F(t) = \mathrm{P}[T^* > t]$ is called the *survival function* of a random variable T^* with a distribution function $F(t)$.

The survival function is a right-continuous non-increasing function as t increases and $S(0) = 1$. It uniquely determines the distribution of the data. Unfortunately, in general, we are not able to easily estimate the survival function from data even though we assume that T^* and C are stochastically independent. Under this condition, the following inequality holds,

$$S_T(t) = \mathrm{P}[T > t] = \mathrm{P}[T^* > t, C > t] = S(t)\mathrm{P}[C > t] \leq S(t),$$

where $S_T(t)$ is the survival function of T. The distribution of C is rarely known thus there is no possibility to derive $S(t)$.

Next, we define another function that is also a unique characterization of the distribution.

Definition. Let T^* be a continuous non-negative random variable. Then the *hazard function* $\lambda(t)$ of T^* is defined as

$$\lambda(t) = \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}[t \leq T^* < t + h | T^* \geq t].$$

Let T^* be discrete with values $0 \leq t_1 < t_2 < \ldots$. Then the *hazard function* $\lambda(t)$ of T^* is defined at $t_2, t_2, \ldots$ by

$$\lambda(t_i) \equiv \lambda_i = \mathrm{P}[T^* = t_i | T^* \geq t_i].$$

In other words, the hazard function $\lambda(t)$ expresses the probability of occurrence of an event in the time interval $[t, t + h)$ assuming that no event for the monitored subject occurred before time t. The useful property of the hazard function is that it can be estimated from censored data under certain conditions. It is much simpler than in the case of the survival function and we do not need to know the distribution of C. The hazard function is therefore a convenient tool to analyze time-to-event data. We then define the cumulative hazard function.

Definition. The function $\Lambda(t)$ defined as

$$\Lambda(t) = \int_0^t \lambda(s) ds,$$

for continuous T^*, and

$$\Lambda(t) = \sum_{i: t_i \leq t} \lambda(t_i) ds,$$

for discrete T^*, is called the *cumulative hazard function.*

It describes the accumulated risk up until time t and it can also be interpreted as the expected number of events to be observed by time t.

Earlier we assumed a stochastic independence of the censoring time C and the survival time T^*, however it is not necessary to insist on such a strong condition. Instead, we impose the independent censoring condition.

Definition. The censoring variable C satisfies *the independent censoring condition* for the failure time T^* with cumulative hazard Λ if and only if

$$\Lambda(t) = -\int_0^t \frac{d\mathrm{P}[T^* \geq s, C \geq T^*]}{\mathrm{P}[T^* \geq s, C \geq s]} \quad \forall s \text{ such that } \mathrm{P}[T^* \geq s, C \geq s] > 0. \tag{1.3}$$

Remark. By expression in the numerator of (1.5) we understand

$$P(t < T^* \leq t + s, T^* \leq C) = -\int_{z=t}^{z=t+s} d\mathrm{P}(T^* > z, C \geq T^*).$$

Remark. Let T^* has a continuous distribution then (1.5) is equivalent to equality,

$$\begin{aligned}\lambda(t) &= \frac{-\frac{\partial}{\partial s}\mathrm{P}[T^* \geq s, C \geq t]|_{s=t}}{\mathrm{P}[T^* \geq t, C \geq t]} \\ &= \lim_{h \searrow 0} \frac{1}{h}\mathrm{P}[t \leq T^* < t + h | T^* \geq t, C \geq t] \quad \forall t \geq 0. \end{aligned} \tag{1.4}$$

We assume that the independent censoring condition holds for further derivations.

1.2.2 The Cox model

We assume $T^* \geq 0$ follows a continuous distribution. We observe K independent triplets $(T_i, \delta_i, \tilde{\mathbf{X}}_i)$, where $T_i = \min(C_i, T_i^*)$, $\delta_i = \mathbb{1}(T_i^* \leq C_i)$ and $\tilde{\mathbf{X}}_i = (\tilde{X}_{i1}, \ldots, \tilde{X}_{ir})^\mathsf{T}$ is a vector of covariates obtained for each subject. The covariates could be fixed (time-independent) or they may be functions of time. We are interested in whether the covariates are significantly influencing the distribution of T_i^*. We are searching for the model that allows us to estimate the effect of the covariates on the risk of having an event.

Including covariates in the model allows us to use a weaker condition of independent censoring. Instead of (1.5) we define an independent censoring condition given the covariates $\tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}}(t)$ is a vector of values of the covariates at time t.

Definition. The censoring variable C satisfies *the independent censoring condition* for the failure time T^* with cumulative hazard Λ given the covariates $\tilde{\mathbf{X}}$ if and only if

$$\begin{aligned}\lambda(t|\tilde{\mathbf{X}}) \equiv \lim_{h \searrow 0} \frac{1}{h}\mathrm{P}[t \leq T^* < t + h | T^* \geq t, \tilde{\mathbf{X}}(t)] = \\ \lim_{h \searrow 0} \frac{1}{h}\mathrm{P}[t \leq T^* < t + h | T^* \geq t, C \geq t, \tilde{\mathbf{X}}(t)] \quad \forall t \geq 0. \end{aligned} \tag{1.5}$$

The sufficient condition for this equality is conditional independence of T and C given the covariates. Thus, the distribution of censoring times may vary among different population groups as long as a covariate distinguishing between

these groups included in the model (e.g. the students of primary school, secondary school and high school can have different censoring distributions when the variable defining the types of education is involved in the model).

We used to define the regression models for the expectation of the response (in our case this stands for the expected failure time). However due to the nature of our data it is more meaningful to build a model for the conditional hazard function. The Cox proportional hazards model (Cox [1972]) is a type of regression model that is suitable for our purposes and it is a common tool that is also used in joint modelling.

Definition. The pairs $(T_i^*, \tilde{\mathbf{X}}_i(t)), i = 1, \ldots, K$ satisfy the *Cox proportional hazards model* if the following two conditions hold:

(i) they are independent across different subjects,

(ii) the conditional hazard function given the covariate process has the form

$$\lambda(t|\tilde{\mathbf{X}}) = \lambda_0(t)\exp\{\boldsymbol{\alpha}_0^\mathsf{T}\tilde{\mathbf{X}}(t)\}, \tag{1.6}$$

where $\lambda_0(t)$ is some unknown unspecified hazard function and $\boldsymbol{\alpha}_0 \in \mathbb{R}^r$ is an unknown vector of regression coefficients.

Then the regression parameters are estimated by the partial likelihood method (Cox [1972]). This method does not require a specification of $\lambda_0(\cdot)$, i.e. a Cox model is a semi-parametric model where $\lambda_0(\cdot)$ is a non-parametric part. The interpretation of the parameters is as follows. The covariates $\tilde{\mathbf{X}}$ act multiplicatively on the conditional hazard function. Suppose that $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{X}}^*$ are the values of covariates for two subjects that do not depend on time. The hazard ratio or relative risk can be expressed with the help of (1.6) as,

$$\frac{\lambda(t|\tilde{\mathbf{X}}^*)}{\lambda(t|\tilde{\mathbf{X}})} = \exp\{\boldsymbol{\alpha}_0^\mathsf{T}(\tilde{\mathbf{X}}^* - \tilde{\mathbf{X}})\} \overset{\tilde{\mathbf{X}}^* = \tilde{\mathbf{X}} + \mathbf{e}_j}{=} \exp\{\alpha_{0j}\},$$

where $\mathbf{e}_j$ is a vector with 1 on j-th place and 0 otherwise. It follows that the relative risk for the event is equal to an exponentiated regression parameter if there is an increase in one covariate by one unit, while other covariates remain fixed. The relative risk is same at each time point and this characteristic is called the proportional hazard assumption.

We are not restricted just to variables whose values are constant during the follow-up period, e.g. gender, type of education, etc. The time varying variables can also be taken into account when we are estimating the risk of an event. The extension of the proportional Cox model allows us to use this type of covariates. However, then the assumption of proportional hazard does not hold any more. The exponentiated regression coefficient then expresses the relative risk at a given time point t when there is a unit increase in corresponding covariate and other covariates remain unchanged. Note that these variables are assumed to be more like piecewise-constant covariates due to the fact that the measurements are provided just for particular time points t_{ij}. The second option is that they may be created as an interaction of time-independent covariate with some function of time $g(t)$.

Without any restrictions we can include the so-called exogenous (external) time varying variables in the model. Suppose that $\tilde{\mathcal{X}}_i(t) = \{\tilde{\mathbf{X}}_i(s), 0 \leq s < t\}$ denote the history of covariates $\tilde{\mathbf{X}}_i$ for subject i up to time t. The exogenous covariates are those which satisfy,

$$\mathrm{P}[\tilde{\mathcal{X}}_i(t)|\tilde{\mathcal{X}}_i(s), T_i \geq s] = \mathrm{P}[\tilde{\mathcal{X}}_i(t)|\tilde{\mathcal{X}}_i(s), T_i = s], \quad 0 < s \leq t. \tag{1.7}$$

In other words, the equality (1.7) formalizes the idea that their future path is not influenced by the occurrence of the event, for instance, weather conditions, humidity etc. On the other hand, endogenous time varying variables are those which are observed repeatedly on the subject and may somehow be associated with the occurrence of failure at the time s, i.e. they are not predictable. Concerning some medical studies, this type of data is often encountered. Moreover, it is reasonable to assume that these observations are contaminated with measurement errors and the whole path is not observed. The extended Cox model requires for variables to be a predictable process with a full path to be specified and measured without error. This is not satisfied here. Thus, modeling these markers using a suitable model before including them into the Cox proportional model appears to be a way to solve this issue and it gave an incentive to the formulation of joint models for longitudinal data and time-to-event data (Rizopoulos [2012], Chapter 4).

1.3 Bayesian statistics

The last part of this introductory chapter focuses on Bayesian statistics. We do not want to introduce a whole field that is wide and extensive. The main purpose of this section is to introduce a notation, terms that are going to be used in later chapters and a short comment about the Gibbs and Metropolis-Hastings algorithm.

1.3.1 Basic concepts

Let $\boldsymbol{Y}$ represent data that comes from probability distribution that depends on the unknown parameter $\boldsymbol{\theta} \in \Theta$, where $\Theta \in \mathbb{R}^p$ is an appropriate parametric space. From a frequentist point of view, the data $\boldsymbol{Y}$ is generated from distribution with a density $p(\boldsymbol{y}|\boldsymbol{\theta})$ with respect to σ-finite measure λ, and the likelihood function of the parameter $\boldsymbol{\theta}$ is defined as $L(\boldsymbol{\theta}) = p(\boldsymbol{Y}|\boldsymbol{\theta})$. The main idea of Bayesian statistical methods is based on the fact that there is some prior knowledge of the distribution of the true value of the parameter $\boldsymbol{\theta}$. This distribution is a so-called *prior distribution* with a corresponding *prior density* $p(\boldsymbol{\theta})$ with respect to σ-finite measure ν.

The idea of how to involve all available information about $\boldsymbol{\theta}$, i.e. observed data $\boldsymbol{y}$ and a prior distribution of $\boldsymbol{\theta}$, to the statistical inference of $\boldsymbol{\theta}$ is provided by the Bayes‘ theorem.

Theorem 1 (Bayes‘ theorem)**.** *A conditional distribution of $\boldsymbol{Y}|\boldsymbol{\theta}$ is determined by a density $p(\boldsymbol{y}|\boldsymbol{\theta})$ and a prior density of a random parameter $\boldsymbol{\theta}$ is $p(\boldsymbol{\theta})$. Then the density of the distribution $\boldsymbol{\theta}|\boldsymbol{Y}$ is of a form*

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) = \begin{cases} \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, & \text{if } p(\boldsymbol{y}) = \int_\Theta p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\nu(\boldsymbol{\theta}) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. See (Anděl [2007], Chapter 3.5, Theorem 3.21). □

The density $p(\boldsymbol{\theta}|\boldsymbol{Y})$ is called *a posterior density* and it determines *a posterior distribution* of $\boldsymbol{\theta}$ under the knowledge of data $\boldsymbol{Y}$. By applying the Bayes' theorem it is possible to obtain the posterior density of $\boldsymbol{\theta}$ that is later used to calculate an estimate of $\boldsymbol{\theta}$. This can be understood as that the estimate of $\boldsymbol{\theta}$ is a kind of update of the prior value of the parameter after taking into account collected data. We denote by $\boldsymbol{\theta}_j \in \mathbb{R}^q,\ q < p$, a subvector of $\boldsymbol{\theta}$. Similarly, by $\boldsymbol{\theta}_{-j}$, we understand a subvector of $\boldsymbol{\theta}$ such that j-th component is missing, i.e. $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_1^\mathsf{T}, \ldots, \boldsymbol{\theta}_{j-1}^\mathsf{T}, \boldsymbol{\theta}_{j+1}^\mathsf{T}, \ldots, \boldsymbol{\theta}_p^\mathsf{T})^\mathsf{T}$.

An important term that is related to posterior distribution is *full conditional distribution of* $\boldsymbol{\theta}_j$. The density of full conditional distribution of $\boldsymbol{\theta}_j$ is denoted as $p(\boldsymbol{\theta}_j|\boldsymbol{y}, \boldsymbol{\theta}_{-j})$ and it is proportional to the likelihood function multiplied by the prior density of the parameter, i.e.,

$$p(\boldsymbol{\theta}_j|\boldsymbol{y}, \boldsymbol{\theta}_{-j}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

In general, it is easier to calculate the full conditional distribution of $\boldsymbol{\theta}_j|\boldsymbol{y}, \boldsymbol{\theta}_{-j}$ compared to *marginal distribution of* $\boldsymbol{\theta}_j|\boldsymbol{y}$ due to the fact that the derivation of the marginal distribution requires us to compute an integral over all components of $\boldsymbol{\theta}_{-j}$. That is not always possible to solve the integral analytically or it can be numerically intensive when the closed form solution does not exist. The density of the marginal distribution of $\boldsymbol{\theta}_j|\boldsymbol{y}$ takes a form,

$$p(\boldsymbol{\theta}_j|\boldsymbol{y}) \propto \int_{\Theta_{-j}} p(\boldsymbol{y}|\boldsymbol{\theta}), p(\boldsymbol{\theta}) \mathrm{d}\mu(\boldsymbol{\theta}_{-j}).$$

This was a short overview of the important terms and a summary of the basic concept of Bayesian statistics. For more information on Bayesian statistics, we recommend the following literature, e.g. (Robert [2007]). In the very last section of this introductory chapter we briefly recall two commonly used algorithms for parameter estimation.

The last note is related to the notation. As we are going to discuss just the distributions with a density with respect to a σ-finite measure in this thesis, we simplify the notation to $\mathrm{d}\boldsymbol{\theta}$ instead of $\mathrm{d}\nu(\boldsymbol{\theta})$, i.e. the distribution of $\boldsymbol{\theta}|\boldsymbol{y}$ is denoted as $p(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{y})$.

1.3.2 Estimation methods

In contrast to the classic statistics where $p(\boldsymbol{Y}|\mathrm{d}\boldsymbol{\theta})$ is employed to estimate an unknown parameter, we would like to engage a prior knowledge about $\boldsymbol{\theta}$ to the estimation process. It follows that the posterior distribution should be involved in the estimation process.

Naturally, a posterior expected value appears to be a suitable form of estimator of $\boldsymbol{\theta}$, i.e. a conditional expected value with respect to a posterior distribution $\boldsymbol{\theta}|\boldsymbol{Y}$,

$$\hat{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{Y}) = \int_\Theta \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{Y}) \mathrm{d}\boldsymbol{\theta}. \tag{1.8}$$

To derive such an estimator, we need to know the posterior distribution, or we have to compute the integral (4.1) at least numerically. We are also often interested in credible intervals or regions as equal-tailed (ET) intervals or the highest posterior density (HPD) regions, quantiles etc. for individual components of $\boldsymbol{\theta}$. However, except for the trivial cases, it requires us to derive individual marginal posterior distributions and then the computation of the integral with respect to the remaining components of $\boldsymbol{\theta}$. That can be tough in a complex model and thus, there is a need for different estimation methods to get an estimate of $\boldsymbol{\theta}$.

On that account, we are shortly going to mention the Markov chains Monte Carlo (MCMC) methods that are used to compute the estimates of parameters of interest in a more efficient way. In general these methods are based on simulations and the goal is to generate a Markov chain with certain properties. In general we are searching for a Markov chain with a limit distribution $p(\mathrm{d}\boldsymbol{\theta})$. A thorough introduction to the theory behind is well described in contemporary literature (Brooks et al. [2011], Chapter 1). Here, we will not define any terms related to the theory of Markov chains. As a reminder, we only present two algorithms that are commonly applied tools in Bayesian statistics and that can be used to compute the estimates of the model discussed in this thesis and to which we refer in the theoretical part of the thesis.

The length of the generated Markov chain is $B + M$, where B settles for a so-called *burn-in period* so the first B states are not used to compute the estimate. The remaining M states are used for posterior inference.

Let start with a Gibbs algorithm (Geman and Geman [1984]). There is a set of assumptions that must be satisfied. A parametric space Θ has to be of a form $\Theta = \prod_{j=1}^{k} \Theta_j, \boldsymbol{\theta} = (\boldsymbol{\theta}_1^\mathsf{T}, \ldots, \boldsymbol{\theta}_k^\mathsf{T})^\mathsf{T}$. A target (stationary) distribution is $p(\mathrm{d}\boldsymbol{\theta})$ with a density $p(\boldsymbol{\theta})$ with respect to a product measure $\lambda_1 \otimes \cdots \otimes \lambda_k$, where λ_j is a σ-finite measure such that $\lambda_j(\Theta_j) > 0, \forall j \in \{1, \ldots, k\}$, next we need that $\Theta = \{\boldsymbol{\theta} : p(\boldsymbol{\theta}) > 0\}$. The last important assumption is that we are able to (easily) generate from full conditional distributions $p(\mathrm{d}\boldsymbol{\theta}_j | \boldsymbol{\theta}_{-j})$, where $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_1^\mathsf{T}, \ldots, \boldsymbol{\theta}_{j-1}^\mathsf{T}, \boldsymbol{\theta}_{j+1}^\mathsf{T}, \ldots, \boldsymbol{\theta}_k^\mathsf{T})^\mathsf{T}$. The algorithm itself is composed out of three steps.

Gibbs algorithm

1. Select an initial state $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)\mathsf{T}}, \ldots, \boldsymbol{\theta}_k^{(0)\mathsf{T}})^\mathsf{T}$, and set $m = 0$,

2. (i) generate $\boldsymbol{\theta}_1^{(m+1)}$ from the conditional distribution

$$p(\mathrm{d}\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(m)}, \ldots, \boldsymbol{\theta}_k^{(m)}, \boldsymbol{y}),$$

(ii) generate $\boldsymbol{\theta}_2^{(m+1)}$ from the conditional distribution

$$p(\mathrm{d}\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(m+1)}, \boldsymbol{\theta}_3^{(m)}, \ldots, \boldsymbol{\theta}_k^{(m)}, \boldsymbol{y}),$$

(iii) generate $\boldsymbol{\theta}_3^{(m+1)}$ from the conditional distribution

$$p(\mathrm{d}\boldsymbol{\theta}_3 | \boldsymbol{\theta}_1^{(m+1)}, \boldsymbol{\theta}_2^{(m+1)}, \boldsymbol{\theta}_4^{(m)} \ldots, \boldsymbol{\theta}_k^{(m)}, \boldsymbol{y}),$$

$\vdots$

(k) generate $\boldsymbol{\theta}_k^{(m+1)}$ from the conditional distribution

$$p(\mathrm{d}\boldsymbol{\theta}_k|\boldsymbol{\theta}_1^{(m+1)},\ldots,\boldsymbol{\theta}_{k-1}^{(m+1)},\boldsymbol{y}).$$

3. Then add one to m and goto step 2.

The output of the algorithm is a Markov chain $\boldsymbol{\theta}^{(m)},\ m\in\{1,\ldots,B+M\}$, where the first B components are not used for the calculation of the estimate of $\boldsymbol{\theta}$ or $t(\boldsymbol{\theta})$, where $t(\cdot)$ is some measurable function. Ergodicity is assured if the assumptions mentioned above are met.

Second, we briefly describe the Metropolis-Hastings algorithm (Metropolis et al. [1953], Hastings [1970]). The assumptions for the application of the algorithm are the following: a target (stationary) distribution is $p(\mathrm{d}\boldsymbol{\theta})$ with a density $p(\boldsymbol{\theta})$ with respect to σ-finite measure λ such that $\lambda(\Theta)>0$, where $\Theta=\{\boldsymbol{\theta}:p(\boldsymbol{\theta})>0\}$ is a parametric space.

Metropolis-Hastings algorithm

1. Select an initial state $\boldsymbol{\theta}^{(0)}$, and set $m=0$,
2. generate a proposal of $\boldsymbol{\psi}$ from a distribution $q(\boldsymbol{\theta}^{(m)},\mathrm{d}\boldsymbol{\psi})$ with a density $q(\boldsymbol{\theta}^{(m)},\boldsymbol{\psi})$ (with respect to σ-finite measure λ),
3. compute the proposal acceptance probability

$$\alpha(\boldsymbol{\theta}^{(m)},\boldsymbol{\psi})=\begin{cases}\min\left\{\frac{p(\boldsymbol{\psi})q(\boldsymbol{\psi},\boldsymbol{\theta}^{(m)})}{p(\boldsymbol{\theta}^{(m)})q(\boldsymbol{\theta}^{(m)},\boldsymbol{\psi})},1\right\} & \text{for } p(\boldsymbol{\theta}^{(m)})q(\boldsymbol{\theta}^{(m)},\boldsymbol{\psi})>0,\\ 1, & \text{otherwise,}\end{cases}$$

4. generate $U\sim\mathcal{U}(0,1)$, where $\mathcal{U}$ stands for a uniform distribution and

$$\boldsymbol{\theta}^{(m+1)}=\begin{cases}\boldsymbol{\psi}, & \text{if } U<\alpha(\boldsymbol{\theta}^{(m)},\boldsymbol{\psi})\\ \boldsymbol{\theta}^{(m)}, & \text{if } U\geq\alpha(\boldsymbol{\theta}^{(m)},\boldsymbol{\psi}),\end{cases}$$

5. then add one to m and goto step 2.

The output of the algorithm is again a Markov chain $\boldsymbol{\theta}^{(m)},\ m\in\{1,\ldots,B+M\}$, where the last M components are used to calculate the estimate of $\boldsymbol{\theta}$.

There is no need to know the normalizing constant of the target density $p(\boldsymbol{\theta})$, so this algorithm is suitable for Bayesian statistics as we often work with the form of density without a normalizing constant. The proposal density $q(\boldsymbol{\theta},\boldsymbol{\psi})$ is allowed to be arbitrary, however, the choice of $q(\boldsymbol{\theta},\boldsymbol{\psi})$ can have a huge impact on the proposal acceptance probability. In addition, to ensure the ergodicity, the proposal density has to satisfy certain conditions. An example of such density is a symmetric random walk. For more details we refer to (Brooks et al. [2011]).

2. Joint models for longitudinal and time-to-event data

The basic idea of joint modelling is to enable analysis of time-to-event data with repeated measurements as a predictor taken into account. It leads to reduction of biases, more precise predictions (Njeru Njagi et al. [2013]), and it improves efficiency of statistical inferences. First, we define the joint distribution of an event time T and a longitudinal marker $\boldsymbol{Y}$.

Let $\boldsymbol{Y}$ be a response vector for all subjects and $\boldsymbol{T}$ is a time-to-event, also for all subjects. We assume that $\boldsymbol{Y}$ and time-to-event $\boldsymbol{T}$ jointly follow a continuous distribution with a density $f_{\boldsymbol{Y},\boldsymbol{T}}(\boldsymbol{y},\boldsymbol{t})$.

In this chapter we focus on one of the possible approaches to specify this density. This approach considers a latent variable $\boldsymbol{U}$ that is assumed to capture dependency between $\boldsymbol{T}$ and $\boldsymbol{Y}$. The latent variable $\boldsymbol{U}$ follows a distribution $F_{\boldsymbol{U}}(\boldsymbol{u})$ and we assume that the conditional independence of $\boldsymbol{Y}$ and $\boldsymbol{T}$ given a latent variable holds (Diggle et al. [2008]). Then the joint distribution of $(\boldsymbol{Y},\boldsymbol{T})$ can be written in a form,

$$f_{\boldsymbol{Y},\boldsymbol{T}}(\boldsymbol{y},\boldsymbol{t}) = \int f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) f_{\boldsymbol{T}|\boldsymbol{U}}(\boldsymbol{t}|\boldsymbol{u}) \mathrm{d}F_{\boldsymbol{U}}(\boldsymbol{u}). \tag{2.1}$$

Especially, the shared dependency could be explained by the shared random effects. A widely used combination is a normal linear model with random effects and a Cox proportional hazard model. Another approach is a joint latent class model, where it is assumed that the longitudinal marker $\boldsymbol{Y}$ and time-to-event T are conditionally independent given some discrete latent variable (Diggle et al. [2008]). Those two methods are shortly introduced in the following sections.

Remark. For simplicity we later use $f(\boldsymbol{y})$ as a notation for the density of $\boldsymbol{Y}$ instead of $f_{\boldsymbol{Y}}(\boldsymbol{y})$.

2.1 Joint models with shared random effects

A widespread type of models for joint modelling where the shared latent structure is specified by random effects is called a shared random effects model (SREM). The longitudinal trajectory is a function of those random effects and relevant predictors. Subsequently, this function is included in a survival model as a predictor. Consequently, it follows from (2.1) that the distribution of the longitudinal marker and time-to-event is assumed to be conditionally independent given the random effects.

A common choice to describe the subject-specific trajectory of longitudinal response for K independent subjects is a linear mixed-effects model (LMEM). The random effects take a role of the latent variable $\boldsymbol{U}$ in (2.1). To be consistent with the notation of the theory of linear mixed-effects models, the letter $\boldsymbol{b}$ is used for random effects (latent variable) instead of $\boldsymbol{U}$.

Suppose that $f(\boldsymbol{b}_i)$ is a density of $\boldsymbol{b}_i$. From the definition of the LME model, the independence of subjects is assumed, thus (2.1) can be written in this special case as a product of densities corresponding to individual subjects, i.e.

$$f(\boldsymbol{y}, \boldsymbol{t}) = \prod_{i=1}^{K} f(\boldsymbol{y}_i, t_i) = \prod_{i=1}^{K} \int f(\boldsymbol{y}_i|\boldsymbol{b}_i) f(t_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i) \mathrm{d}\boldsymbol{b}_i. \tag{2.2}$$

2.1.1 A definition of the model

Assume that the longitudinal response $\boldsymbol{Y}$ follows a normal distribution. We rewrite a definition of the linear mixed-effects model (1.1) from the matrix notation by the single values at time t due to later application in a survival model. The observations are split into the true unobserved value $m_i(\cdot)$ of longitudinal marker and the error term $\varepsilon_i(\cdot)$,

$$\begin{aligned} \boldsymbol{Y}_i(t) &= m_i(t) + \varepsilon_i(t), \\ m_i(t) &= \mathbf{X}_i(t)^\mathsf{T}\boldsymbol{\beta} + \mathbf{Z}_i(t)^\mathsf{T}\boldsymbol{b}_i, \end{aligned} \tag{2.3}$$

where the error terms $\varepsilon_i(t) \sim \mathsf{N}(0, \sigma^2)$, the random effects $\boldsymbol{b}_i \sim \mathsf{N}_q(\mathbf{0}, \mathbb{D})$.

Although we only observe data at a few time points we need to estimate the true unobserved value $m_i(\cdot)$ and reconstruct the complete longitudinal history of the marker for each subject so that the longitudinal outcome can be involved in the survival model (Rizopoulos [2012], Chapter 4). To quantify the strength of the relationship between the risk for the event, the longitudinal marker and other factors it is reasonable to postulate a Cox proportional hazards model of the following form,

$$\lambda_i(t|\boldsymbol{\zeta}, \boldsymbol{\gamma}, \alpha) = \lambda_0(t; \boldsymbol{\zeta}) \exp\{\tilde{\mathbf{X}}_i(t)^\mathsf{T}\boldsymbol{\gamma} + \alpha m_i(t)\}, \tag{2.4}$$

where $\lambda_0(t; \boldsymbol{\zeta})$ is an unspecified baseline risk function, $\tilde{\mathbf{X}}_i(t)$ is an r-vector of values of covariates at time t. Commonly, some components of $\tilde{\mathbf{X}}_i(t)$ coincide with the components of $\mathbf{X}_i(t)$. Next, $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{r-1})^\mathsf{T}$ is a vector of unknown parameters and α is an unknown parameter capturing the association between the risk of an event and the true value of the marker $m_i(t)$ at time t. The hypothesis $H_0 : \ \alpha = 0$ indicates that we are testing whether there is a significant effect of the current value of the longitudinal outcome on the risk of the event.

Remark. The risk of an event does not have to be associated only with the current value of the marker. Obviously, there are several possibilities of how to specify an association structure due to the nature of the longitudinal marker, e.g. lagged value of marker, cumulative value of marker, or an interaction with other covariates. The generalized version of (2.4) is,

$$\lambda_i(t|\boldsymbol{\zeta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \lambda_0(t; \boldsymbol{\zeta}) \exp\{\tilde{\mathbf{X}}_i(t)^\mathsf{T}\boldsymbol{\gamma} + h(m_i(t-c), \boldsymbol{\alpha}, \tilde{\mathbf{S}}_i(t))\}, \tag{2.5}$$

where $h(\cdot)$ is a function of the true level of the longitudinal marker $m_i(\cdot)$, of a set of covariates $\tilde{\mathbf{S}}_i(\cdot)$, and the vector of unknown parameters $\boldsymbol{\alpha}$. More information about the several forms of parametrizations of the longitudinal outcome is covered by (Rizopoulos [2012]).

A standard method used to estimate the parameters is a combination of the maximal likelihood estimation and EM algorithm. The second option is a Bayesian approach that may be advantageous at some point. It often leads to easier computational implementation and it allows us to address complex models that involve multivariate longitudinal outcomes or survival observations (Lawrence Gould et al. [2015]). However, the majority of literature focuses on univariate joint modelling, thus just one time dependent covariate is included in the survival model. The extension of univariate joint modelling to the investigation of multivariate longitudinal outcomes results in high dimensionality of random effects and a large number of parameters. For instance, this is often the case in medical studies where multiple biomarkers are repeatedly recorded on a patient and we want to include them all in a single joint model (Hickey et al. [2016]). That is computationally more intensive even if the Bayesian approach is used. An alternative to SREM is a joint latent class model where the main interest is in prediction and we are not focused on the interpretation of the parameters (Proust-Lima et al. [2014]).

2.2 Joint latent class models

An alternative approach to the shared random-effect model is the joint latent class model (JLCM). The basic principle is based on the fact that the dependency between the event time and longitudinal predictor can be described by a latent class structure which is not observable.

Suppose that the population is not homogeneous and assume that there is a finite number of heterogeneous groups in the population. The common characteristics within the group are the same risk of the event and the same trajectory of the time dependent variable, i.e. the class structure captures the entire relationship between the longitudinal marker and the time-to-event. Thus the subject specific trajectory of the longitudinal marker and the risk of an event can be assumed to be conditionally independent (Proust-Lima et al. [2014]). Let V_i is a categorical latent variable with G possible values, then (2.1) can be rewritten as,

$$f(\boldsymbol{y}_i, t_i) = \sum_{g=1}^{G} f(\boldsymbol{y}_i | V_i = g) f(t_i | V_i = g) \mathrm{P}(V_i = g). \tag{2.6}$$

2.2.1 A definition of the model

Suppose $i = 1, \ldots, K$ is a number of independent subjects, $g = 1, \ldots, G$ is a number of classes, π_{ig} is a probability that a subject i belongs to the latent class g. The probabilities of class membership can be defined in several ways and it is also allowed that they depend on subject specific covariates. One of the commonly used options is to define probabilities through subject-specific covariates is,

$$\pi_{ig} = \mathrm{P}[V_i = g | \tilde{\mathbf{Z}}_i] = \frac{\exp\{\tilde{\mathbf{Z}}_i^{\mathsf{T}} \boldsymbol{\xi}_g\}}{\sum_{l=1}^{G} \exp\{\tilde{\mathbf{Z}}_i^{\mathsf{T}} \boldsymbol{\xi}_l\}}, \tag{2.7}$$

where $\tilde{\mathbf{Z}}_i$ is an m-vector of time-independent covariates for a subject i, $\boldsymbol{\xi}_g = (\xi_{0g}, \ldots, \xi_{m-1,g})^{\mathsf{T}}, g = 1, \ldots, G$ are the unknown parameters. To have all parameters identifiable, it is assumed $\boldsymbol{\xi}_G = \mathbf{0}$.

First, we need to build models for the longitudinal marker and the risk of the event, respectively. The natural choice of the model for the marker again is a linear mixed effects model. For this model, given V_i, it holds that $\boldsymbol{Y}_i, \boldsymbol{b}_i|_{V_i=g} = \boldsymbol{Y}_i|_{\boldsymbol{b}_i,V_i=g} \times \boldsymbol{b}_i|_{V_i=g}$, where $\boldsymbol{Y}_i|_{\boldsymbol{b}_i,V_i=g}$ takes a well-known form,

$$\boldsymbol{Y}_i|_{\boldsymbol{b}_i,V_i=g} = \mathbb{X}_i\boldsymbol{\beta}_g + \mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon_i}, \tag{2.8}$$

where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\mathsf{T}$ is a response vector for i-the subject, $\mathbb{X}_i$ is an $n_i \times p$ design matrix. The covariates included in $\mathbb{X}_i$ can differ compared to covariates employed in (2.7) due to the fact that the time-dependent covariates are allowed. A p-vector $\boldsymbol{\beta}_g = (\beta_{0g}, \ldots \beta_{p-1,g})^\mathsf{T}$ is the vector of unknown regression parameters for g-th class. Let $\mathbb{Z}_i$ be an $n_i \times q$ design matrix for random effects $\boldsymbol{b}_i$. The marginal distribution of random effects is a mixture of normal distributions where the weights correspond to the probabilities π_{ig} given by (2.7), i.e. $\boldsymbol{b}_i \sim \sum_{g=1}^{G} \pi_{ig}\mathsf{N}_q(\boldsymbol{\mu}_g, \mathbb{D}_g)$ or equivalently given the class $\boldsymbol{b}_i|_{V_i=g} \sim \mathsf{N}_q(\boldsymbol{\mu}_g, \mathbb{D}_g)$. An n_i-vector of random errors $\boldsymbol{\varepsilon}$ follows $\mathsf{N}_{n_i}(\boldsymbol{0}, \Sigma_i)$. There is no restriction on covariance matrices Σ_i and $\mathbb{D}_g$, however the common choice is a diagonal matrix $\sigma^2\mathbb{I}_{n_i}$ for Σ_i and $\mathbb{D}_g = \omega_g^2\mathbb{D}$ with $\omega_G^2 = 1$.

Second, we have to specify a model for time-to-event. The natural choice of modelling a risk of occurrence of an event is again the Cox proportional hazard model, i.e.

$$\lambda_i(t|V_i = g; \boldsymbol{\zeta}_g, \boldsymbol{\alpha}_g) = \lambda_{0g}(t; \boldsymbol{\zeta}_g)\exp\{\tilde{\mathbf{X}}_i(t)^\mathsf{T}\boldsymbol{\alpha}_g\}, \tag{2.9}$$

where λ_{0g} is a baseline hazard for g-th class, $\tilde{\mathbf{X}}_i(t)$ is a vector of covariates at time t and $\boldsymbol{\alpha}_g = (\alpha_{1g}, \ldots, \alpha_{rg})^\mathsf{T}$ is an r-vector of unknown parameters. Commonly, the partial likelihood function is used to estimate the parameters in the Cox model. The advantage of the partial likelihood function is that the specification of the baseline hazard is not required. Unfortunately, this approach cannot be applied to the JLCM, so the baseline hazard has to be parametrized, e.g. it follows a Weibull distribution or it is piecewise constant (Proust-Lima et al. [2014]).

The assumption of conditional independence allows us to model trajectories in a survival model in a way that they may vary a lot compared to the joint models with shared random effects. The shared random effects are quite restrictive because they act in both a survival and a longitudinal model. Due to this common structure, great flexibility is not allowed (Rizopoulos [2012], Chapter 4). On the other hand, the interpretation of coefficients in JLCM is not that straightforward, so this modeling is recommended to use when someone is interested in prediction more than in explanation of the relationship between a longitudinal marker and the occurrence of en event (Proust-Lima et al. [2014]).

The main advantage when multivariate outcomes are included in the model is that the JLCM keeps the dimension of the random effects low compared to the SREM. Therefore it is less computationally demanding. The standard way to estimate the parameters is to use a maximum likelihood estimation. The likelihood function is under this model much simpler than under the joint model with shared random effects. The optimal number of the latent groups in a population is often determined by the information criteria based on the penalized likelihood with the Bayesian Information Criterion (BIC) being the most common choice.

Note that we mention just a simple case with one longitudinal marker, but the general version with multivariate longitudinal markers is also possible. (Proust-Lima et al. [2013])

In the following chapters we expand on this model, mainly the estimation methods. The main focus is on the Bayesian approach however the frequentist approach is also briefly mentioned.

3. Estimation of parameters in JLCM

When it comes to the JLCM there are two well-known main approaches of how to estimate the unknown parameters in the model. The frequentist approach is comprised of a widely used maximum likelihood method (MLE). Nowadays, this method is definitely employed more frequently when there is a need to fit the models. That correspond to more developed literature and software implementation for this approach (Proust-Lima et al. [2015], Proust-Lima et al. [2014]). Therefore, MLE is just shortly mentioned and no details are discussed. The majority of this chapter is devoted to the Bayesian approach. The definition of the model in Bayesian context, the choice of prior distributions, the discussion of posterior distributions and a selection of the number of classes that are presented. A separate chapter is devoted to the derivation of the full conditional distributions of the individual parameters of interest.

3.1 The maximum likelihood method

The maximum likelihood method is a widely used method that is used in real-life applications when we are interested in estimating the parameters in a model. This method requires a specification of the distribution for the data and is based on the maximization of the likelihood function, i.e. a joint distribution of the sample as a function of the parameters with the random variables fixed at the observed values.

Likelihood function $L_{obs}(\boldsymbol{\theta})$ under the JLCM takes a form,

$$\begin{aligned} L_{obs}(\boldsymbol{\theta}) &= \prod_{n=1}^{K} f(\boldsymbol{y}_i, t_i, \delta_i, \boldsymbol{\theta}) \\ &= \prod_{n=1}^{K} \{\sum_{g=1}^{G} f(\boldsymbol{y}_i, t_i, \delta_i | V_i = g, \boldsymbol{\theta}) \mathrm{P}[V_i = g | \boldsymbol{\theta}_p]\} \\ &= \prod_{n=1}^{K} \Big\{ \sum_{g=1}^{G} f(\boldsymbol{y}_i | V_i = g, \boldsymbol{\theta}_l) \lambda_i(t_i | V_i = g, \boldsymbol{\theta}_s)^{\delta_i} \\ &\quad S(t_i | V_i = g, \boldsymbol{\theta}_s) \mathrm{P}[V_i = g | \boldsymbol{\theta}_p] \Big\}, \end{aligned} \tag{3.1}$$

where the second equality holds due to the assumed conditional independence of longitudinal model and survival model given the class-membership. The vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_p^\mathsf{T}, \boldsymbol{\theta}_l^\mathsf{T}, \boldsymbol{\theta}_s^\mathsf{T})^\mathsf{T}$ can be split into the subvectors of parameters that corresponds to separate parts of the model, i.e. $\boldsymbol{\theta}_l$ includes all parameters from the longitudinal model, $\boldsymbol{\theta}_s$ covers the set of parameters of the survival model and $\boldsymbol{\theta}_p$ corresponds to the parameters that occur in class probabilities.

Based on (3.1) and the definition of the model in section 2.2 the contribution of the longitudinal model to the likelihood function for the i-th subject in the g-th group is

$$f(\boldsymbol{y}_i|V_i=g,\boldsymbol{\theta}_l)=\int f(\boldsymbol{y}_i|\mathbb{X}_i\boldsymbol{\beta}_g+\mathbb{Z}_i\boldsymbol{b}_{gi},V_i=g)f(\boldsymbol{b}_i|\boldsymbol{\mu}_g,\mathbb{D}_g)d\boldsymbol{b}_i$$
$$=\int\Big(\frac{1}{2\pi\sigma_e^2}\Big)^{\frac{n_i}{2}}\exp\Big\{-\frac{1}{2\sigma_e^2}(\boldsymbol{y}_i-\mathbb{X}_i\boldsymbol{\beta}_g-\mathbb{Z}_i\boldsymbol{b}_i)^\mathsf{T}(\boldsymbol{y}_i-\mathbb{X}_i\boldsymbol{\beta}_g-\mathbb{Z}_i\boldsymbol{b}_i)\Big\}$$
$$\Big(\frac{1}{2\pi}\Big)^{\frac{q}{2}}|\mathbb{D}_g|^{-\frac{1}{2}}\exp\Big\{-\frac{1}{2}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)^\mathsf{T}\mathbb{D}_g^{-1}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)\Big\}d\boldsymbol{b}_i.$$

Next, the contribution of the survival model to the likelihood function for the i-th subject in the g-th group is

$$\lambda_i(t_i|V_i=g,\boldsymbol{\theta}_s)^{\delta_i}S(t_i|V_i=g,\boldsymbol{\theta}_s)=$$
$$=\Big\{\lambda_{0g}(t_i,\nu_g,\eta_g)\exp(\tilde{\mathbf{X}}_i(t)^\mathsf{T}\boldsymbol{\alpha}_g)\Big\}^{\delta_i}\exp(-\Lambda_i(t_i|\nu_g,\eta_g,\boldsymbol{\alpha}_g,\tilde{\mathbf{X}}_i(t))=(\star).$$

Now suppose that $\lambda_{0g}(t)$ follows a Weibull distribution. Moreover, we propose an additional assumption that $\tilde{\mathbf{X}}_i(t)=\tilde{\mathbf{X}}_i$, i.e. covariates are time independent, they are constant over time. This assumption simplifies a function and an integral can be easily calculated,

$$(\star)=\Big\{\nu_g\eta_g t_i^{\nu_g-1}\exp(\tilde{\mathbf{X}}_i(t)^\mathsf{T}\boldsymbol{\alpha}_g)\Big\}^{\delta_i}\exp(-\nu_g\eta_g\int_0^{t_i}s^{\nu_g-1}e^{\tilde{\mathbf{X}}_i(s)^\mathsf{T}\boldsymbol{\alpha}_g}ds)$$
$$=\Big\{\nu_g\eta_g t_i^{\nu_g-1}\exp(\tilde{\mathbf{X}}_i^\mathsf{T}\boldsymbol{\alpha}_g)\Big\}^{\delta_i}\exp(-\eta_g {t_i}^{\nu_g}e^{\tilde{\mathbf{X}}_i^\mathsf{T}\boldsymbol{\alpha}_g}),$$

where the equality $(\star)$ holds under the assumption that $\lambda_{0g}(t)$ follows a Weibull distribution. The last equality $(\star\star)$ holds if $\tilde{\mathbf{X}}_i(t)=\tilde{\mathbf{X}}_i$, i.e. covariates are time independent, they are constant over time.

It is obvious from the form of likelihood that this would not be easy or that for some parameters it is even impossible to calculate estimates as a maximum of log-likelihood analytically. There is a number of algorithms that can be used to maximize the log-likelihood numerically e.g. algorithms in the Newton-Raphson family or in the EM family. Based on satisfactory results from previous analyzes, an extended iterative Marquardt algorithm is used because of its speed and the rate of convergence. In the context of JLCM, this algorithm has already been implemented in R in the `lcmm` package and is used by `Jointlcmm` and `jlcmm` (Proust-Lima et al. [2015]). Since we are forced to use numerical procedures, it is strongly recommended to carry out the estimation process several times to achieve a global maximum instead of a local maximum or to establish initial values that will help us reach a global maximum. (Proust-Lima et al. [2015]). Before starting the procedure, the user must decide on the number of classes in the population. This is often determined using the Bayesian Information Criterion (BIC), nevertheless it is not the only option (Han et al. [2007]). The MLE is not going to be discussed here in more detail because it was already covered in recent years by the available literature (e.g. Proust-Lima et al. [2015], Rizopoulos [2012], Chapter 4).

3.2 The Bayesian approach

We proceed from the model defined in the previous section 2.2 by (Proust-Lima et al. [2014]), however we will look at the model from the perspective of Bayesian

statistics. To our best knowledge, the focus on the application of the Bayesian statistics for this type of model is not that common and not many details about this approach are available. In contemporary literature a model combining JLCM and SREM together is proposed and estimated using Bayesian statistics (Andrinopoulou et al. [2018]). Their model can be easily transformed into the context of JLCM, by setting one parameter equal to zero. However the theoretical derivations are missing, the authors only introduce the form of the likelihood function and they briefly discuss prior distributions. They also do not specify a formula of the class probabilities, i.e. instead of 2.7 they simply consider $\boldsymbol{\pi}$ that does not depend on any additional parameters. There is no proper discussion about how to estimate the parameters of the model, derivation of full conditional distributions that are needed when using the Gibbs algorithm, etc. Next chapter follows up on the obtained results and complete these missing derivations. The main motivation to derive full conditional distributions results from the use of these distributions in algorithms used to estimate model parameters. For the Gibbs algorithm, the Metropolis-Hastings algorithm or the Metropolis within Gibbs algorithm it is sufficient to generate data from these distributions to achieve a result.

Now we will properly define the model we are working with on the following pages in a Bayesian framework. The basic notation of parameters and random variables remain the same as in Chapter 2. The model is a bit extended, we deduce a full conditional distributions for both options of the above class probabilities, distinguishing the subject-specific probability (each subject has its own probabilities to be a member of the class) and a general one (probabilities are the same for all subjects involved in the study). We work in a setting where a baseline hazard $\lambda_0(t)$ follows the Weibull distribution, however other possibilities exists e.g. piecewise constant, usage of B-splines.

Remember, there are G different groups in the population, a membership in the group is not known in advance and the information about the group membership is included in the model by the latent variable V. The joint density of the model is of almost the same form as (3.1), however the sum over G is missing,

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{V}, \boldsymbol{b}, \boldsymbol{\theta}) &= \prod_{n=1}^{K} p(\boldsymbol{y}_i, t_i, \delta_i, V_i, \boldsymbol{b}_i | \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= \prod_{n=1}^{K} p(\boldsymbol{y}_i, t_i, \delta_i, \boldsymbol{b}_i | V_i = g, \boldsymbol{\theta}) \mathrm{P}[V_i = g | \boldsymbol{\theta}] p(\boldsymbol{\theta}) \\ &= \prod_{n=1}^{K} p(\boldsymbol{y}_i | V_i = g, \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i | \boldsymbol{\theta}) \lambda_i(t_i | V_i = g, \boldsymbol{\theta})^{\delta_i} \\ &\quad S(t_i | V_i = g, \boldsymbol{\theta}_s) \mathrm{P}[V_i = g | \boldsymbol{\theta}_p] p(\boldsymbol{\theta}). \end{aligned} \tag{3.2}$$

The individual parts of (3.2) are specified in the next chapter after we specify the model in more detail. This is done on the following lines. The Bayesian model specification is,

Observed data

- Longitudinal model: A longitudinal marker $\mathbf{Y}_i|_{V_i=g,\boldsymbol{b}_i} \sim \mathsf{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta}_g + \mathbb{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_{ig})$, where $\boldsymbol{\Sigma}_{ig}$ is an i-th block of block-diagonal matrix $\boldsymbol{\Sigma}_g$, such that $\Sigma_{ig} = \mathbb{Z}_i\mathbb{D}_g\mathbb{Z}_i^{\mathsf{T}} + \sigma_e^2\mathbb{I}_{n_i}$, where $g = 1, \ldots, G$ and $i = 1, \ldots, K$,

- Survival model: A time-to-event $T_i = \min(T_i^*, C_i)$ follows a continuous distribution, $\delta_i = \mathbb{1}(T_i^* \leq C_i)$ is the event indicator. They follow a Cox model with a baseline hazard $\lambda_{0g}(t)$ following the Weibull distribution with a scale parameter η_g and a shape parameter ν_g.

Latent data

- Random effects: $\boldsymbol{b}_i|_{V_i=g} \sim \mathsf{N}_q(\boldsymbol{\mu}_g, \mathbb{D}_g)$ independent for $g = 1, \ldots, G$,
- Class membership: V_i independent with a discrete distribution $\mathcal{A}(\boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iG})^\mathsf{T}$ and $\pi_{ig} = P[V_i = g]$ is the class-membership probability such that $\sum_{g=1}^{G} \pi_{ig} = 1$. As a special example, π_{ig} are allowed to depend on the subject specific covariates $\tilde{\mathbf{Z}}_i$ through a multinomial logistic regression as it is defined in (2.7), i.e.

$$\mathrm{P}[V_i = g|\boldsymbol{\xi}] = \frac{\exp\{\tilde{\mathbf{Z}}_i^\mathsf{T} \boldsymbol{\xi}_g\}}{\sum_{l=1}^{G} \exp\{\tilde{\mathbf{Z}}_i^\mathsf{T} \boldsymbol{\xi}_l\}} \overset{(\star)}{=} \frac{\exp(\xi_g)}{\sum_{l=1}^{G} \exp(\xi_l)}, \tag{3.3}$$

  where the second equality $(\star)$ holds when it is assumed that the probability does not depend on subject specific covariates $\tilde{\mathbf{Z}}_i$ or it can be interpreted as a model just with an intercept, where $\tilde{\mathbf{Z}}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\mathsf{T}$, with 1 on the g-th position. If it does not depend on the individual subject, then we write $\pi_{ig} = \pi_g$. A prior distribution for parameters of the model has to be specified depending on the parametrization, two options arise in our case, we need to choose a prior distribution for $\boldsymbol{\xi}$ or a prior distribution for $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iG})^\mathsf{T}$ itself is specified.

Parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_l, \boldsymbol{\theta}_s)^\mathsf{T}$ are so-called "genuine" parameters, where

- $\boldsymbol{\theta}_p = (\boldsymbol{\xi}_1^\mathsf{T}, \ldots, \boldsymbol{\xi}_G^\mathsf{T})^\mathsf{T}$, where $\boldsymbol{\xi}_g = (\xi_{0g}, \ldots, \xi_{m-1,g})^\mathsf{T}, g = 1, \ldots, G$, is a set of parameters corresponding to the model of group probabilities,
- $\boldsymbol{\theta}_l = (\sigma_e, \boldsymbol{\beta}_1^\mathsf{T}, \ldots, \boldsymbol{\beta}_G^\mathsf{T}, \boldsymbol{\mu}_1^\mathsf{T}, \ldots, \boldsymbol{\mu}_G^\mathsf{T}, \mathrm{vec}(\mathbb{D}_1), \ldots \mathrm{vec}(\mathbb{D}_g))^\mathsf{T}$ is a set of parameters of the longitudinal model,
- $\boldsymbol{\theta}_s = (\nu_1, \ldots, \nu_G, \eta_1, \ldots, \eta_G, \boldsymbol{\alpha}_1^\mathsf{T}, \ldots, \boldsymbol{\alpha}_G^\mathsf{T})^\mathsf{T}$ is a set of parameters of the survival model, where ν_g and η_g are parameters of the baseline hazard and $\boldsymbol{\alpha}_g$ are parameters that capture the relationship between covariates and a risk of event.

Next step is to specify prior distributions of these parameters.

3.2.1 Prior distributions

A necessary part of the definition of the Bayesian model is the specification of prior distributions for model parameters. We deal with a vector of parameters $\boldsymbol{\theta}$ from (3.2). We assume a prior independence of the components of $\boldsymbol{\theta}$, i.e, $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_l)p(\boldsymbol{\theta}_s)p(\boldsymbol{\theta}_p)$. The parameters that depend on class membership are denoted by index g. In the following text the prior distributions for parameters of a longitudinal part of the model, the survival part of the model and class probability model are discussed separately.

The longitudinal model The prior distributions for the longitudinal model are specified i.e., a prior distribution of the parameter vector $\boldsymbol{\theta}_l$ in which the following prior independence is assumed. For class g we obtain, $p(\boldsymbol{\theta}_{lg}) = p(\sigma^2)p(\boldsymbol{\beta}_g)$ $p(\boldsymbol{\mu}_g)p(\mathbb{D}_g)$.

The variance σ^2 is independent of class specification by definition of the model. Commonly, it is used an inverse variance $\tau = \sigma^{-2}$ in Bayesian statistics then the natural choice of the prior distribution is

$$p(\tau) \sim \Gamma(a_\tau, b_\tau), \tag{3.4}$$

where both parameters a_τ and b_τ are positive. As long as not that much information is available we would rather decide to use less informative prior distributions. This corresponds with the choice of $a_\tau \in (0, 1]$ and b_τ close to zero. Nevertheless the choice of b_τ can have a huge impact on the shape of the distribution so it is not rare that this parameter is allowed to be random following a gamma distribution with fixed parameters. Another suitable choice of prior distribution is $p(\tau) \propto \frac{1}{\tau}, \tau > 0$ that is not informative. In latter derivation in the following chapter we assume that (3.4) holds and the parameters a_τ, b_τ are fixed.

As prior distribution for $\boldsymbol{\beta}$ we use a standard option applied in linear models, i.e.

$$p(\boldsymbol{\beta}_g) \sim \mathsf{N}_p(\boldsymbol{\mu}_{\beta_g}, \Sigma_{\beta_g}), \quad \text{where } \Sigma_{\beta_g} = \operatorname{diag}(\sigma^2_{\beta_g 1}, \ldots, \sigma^2_{\beta_g p}), \tag{3.5}$$

assuming that Σ_{β_g} is a symmetric positive definite matrix. Ordinarily, the diagonal components are chosen large to ensure less impact of the prior distribution to posterior inference, an option for $\boldsymbol{\mu}_{\beta_g}$ is $\mathbf{0}$ except the absolute term of the model (β_{0g}). Moreover, $p(\boldsymbol{\beta}_g) \propto \mathbf{1}$ represents a non-informative prior distribution. In case of random hyperparameters, gamma distribution is used for diagonal components of a covariance matrix and a multivariate normal distribution for the mean vector.

The same discussion as above can be provided for the prior distribution of the vector of expected values $\boldsymbol{\mu}_g$ of random effects,

$$p(\boldsymbol{\mu}_g) \sim \mathsf{N}_q(\boldsymbol{\mu}_{0g}, \Sigma_{0g}), \quad \text{where } \Sigma_{0g} = \operatorname{diag}(\sigma^2_{0g1}, \ldots, \sigma^2_{0gq}), \tag{3.6}$$

assuming that Σ_{0g} is a symmetric positive definite matrix.

The last random parameter in the longitudinal model is a covariance matrix $\mathbb{D}_g$. A prior distribution is defined for inverse of $\mathbb{D}_g$. This inverse exists as long as $\mathbb{D}_g > 0$, it holds because $\mathbb{D}_g$ is a covariance matrix and is defined in this way. A suitable prior distribution of $\mathbb{D}_g^{-1}$ is a q-dimensional Wishart distribution,

$$p(\mathbb{D}_g^{-1}) \sim \mathsf{W}_q(d_{\mathbb{D}_g}, \mathbb{B}_{\mathbb{D}_g}), \tag{3.7}$$

where $d_{\mathbb{D}_g}$ are degrees of freedom and a choice $d_{\mathbb{D}_g} \in (q-1, q]$ leads to a weakly informative prior. The matrix $\mathbb{B}_{\mathbb{D}_g}$ is usually chosen as a diagonal matrix with the diagonal elements being again random or large. The large values lead to a weakly informative prior distribution.

The survival model Prior distributions for the parameters of a survival model are presented hereafter. Again a prior independence of components of a parameter

vector $\boldsymbol{\theta}_s$ is assumed, i.e. for class g we obtain $p(\boldsymbol{\theta}_{sg}) = p(\boldsymbol{\alpha}_g)p(\eta_g)p(\nu_g)$. We start with the parameters related to baseline hazard. By the definition of the model, ν_g and η_g must be positive, thus a natural choice of prior distribution for both parameters is gamma distribution, i.e.

$$p(\eta_g) \sim \Gamma(a_{\eta_g}, b_{\eta_g}), \quad p(\nu_g) \sim \Gamma(a_{\nu_g}, b_{\nu_g}), \tag{3.8}$$

where $a_{\eta_g}, b_{\eta_g}, a_{\nu_g}$, and b_{ν_g} are positive. We dealt with this case when we talked about the prior distribution of τ. The same recommendations can be followed when selecting parameter values or non-informative prior distributions, i.e. $p(\nu_g) \propto \frac{1}{\nu_g}, \nu_g > 0$ and $p(\eta_g) \propto \frac{1}{\eta_g}, \eta_g > 0$.

The size of the effects of covariates in a g-th class is captured through an r-dimensional parameter $\boldsymbol{\alpha}_g$. A standard option of prior distribution is multivariate normal distribution with independence within components of the parameter, i.e.

$$p(\boldsymbol{\alpha}_g) \sim \mathsf{N}_r(\boldsymbol{\mu}_{\alpha_g}, \Sigma_{\alpha_g}), \quad \text{where } \Sigma_{\alpha_g} = \text{diag}(\sigma^2_{\alpha_g 1}, \ldots, \sigma^2_{\alpha_g r}), \tag{3.9}$$

assuming that Σ_α is a symmetric positive definite matrix. Diagonal components Σ_{α_g} are selected as large if less effect on the posterior distribution is required. Similarly to the effects in the longitudinal model we define $p(\boldsymbol{\alpha}_g) \propto \mathbf{1}$ as a non-informative prior.

The class probability model Prior distribution of a parameter vector $\boldsymbol{\theta}_p$ in the class probability model depends on the definition of class probabilities. First, if $\boldsymbol{\theta}_p = \boldsymbol{\pi}$ does not depend on any other parameters than a widely used prior distribution is Dirichlet distribution because it is the conjugate prior of multinomial distribution,

$$p(\boldsymbol{\pi}) \sim \mathsf{Dir}(\boldsymbol{a}), \text{ where } \boldsymbol{a} = (a_1, \ldots, a_G)^\mathsf{T}. \tag{3.10}$$

Without prior knowledge about the distribution of classes in the population, it is recommended to use $a = a_1 = \cdots = a_G$. An often choice of a fixed parameter a are e.g. $a = 1$, $p(\boldsymbol{\pi})$ yields to a Uniform prior (Laplace's prior), for $a = 1/2$ we get a so-called Jefrrey's prior (Alvares et al. [2018]). Another suggestion proposed in (Nasserinejad et al. [2017]) is to assume $a < d/2$, where d is the number of class-specific parameters.

Second, the class probability can depend on covariates and is defined as

$$\mathrm{P}[V_i = g|\boldsymbol{\theta}_p] = \frac{\exp\{\tilde{\mathbf{Z}}_i^\mathsf{T}\boldsymbol{\xi}_g\}}{\sum_{l=1}^G \exp\{\tilde{\mathbf{Z}}_i^\mathsf{T}\boldsymbol{\xi}_l\}},$$

where $\boldsymbol{\theta}_p = \boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_G)^\mathsf{T}$. The parameter $\boldsymbol{\xi}$ takes the same role as parameters $\boldsymbol{\alpha}_g$ or $\boldsymbol{\beta}_g$, thus a natural choice of the prior distribution is again a multivariate normal distribution. The dimension depends on the fact whether probabilities are subject-specific or not,

$$p(\boldsymbol{\xi}) \sim \mathsf{N}_{dim}(\boldsymbol{\mu}_{\boldsymbol{\xi}}, \Sigma_{\boldsymbol{\xi}}), \quad \text{where } \Sigma_{\boldsymbol{\xi}} = \text{diag}(\sigma^2_{\boldsymbol{\xi}_1}, \ldots, \sigma^2_{\boldsymbol{\xi}_G}), \tag{3.11}$$

where $dim = G$ when we assume that the class probability does not depend on covariates, and $dim = mG$ with m being a rank of design matrix $\tilde{\mathbb{Z}}$. The prior distribution is weakly informative if $\sigma^2_{\boldsymbol{\xi}_g}$ are large. An option $p(\boldsymbol{\xi}) \propto \mathbf{1}$ yields to a non-informative prior distribution.

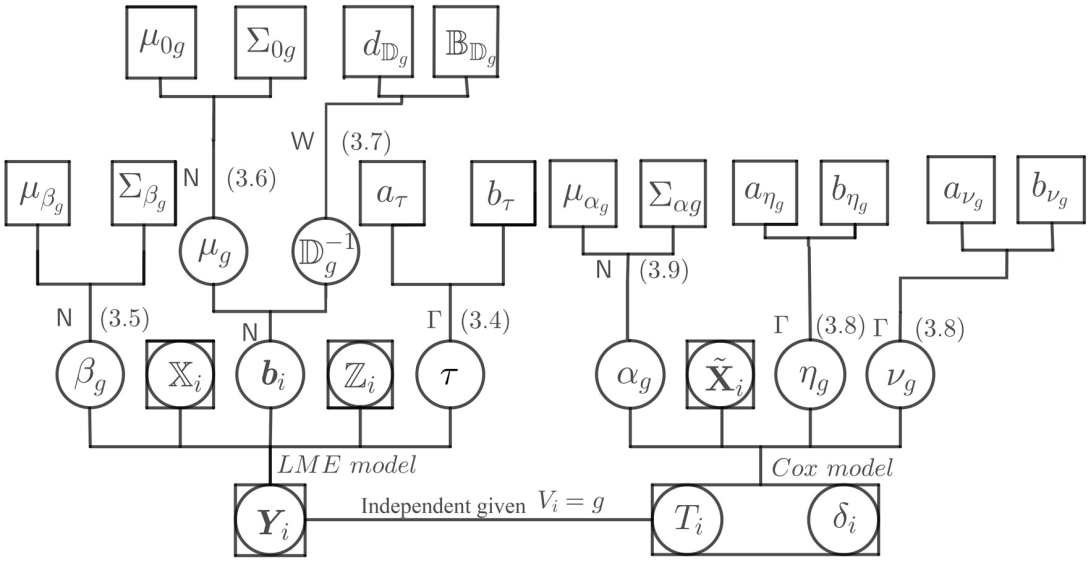

Figure 3.1: Assuming a prior independence, the specification of prior distributions of the parameters of the model given a class g is encoded in a directed acyclic graph (DAG). The numbers in brackets corresponds to the equations.

Remark. A not negligible number of hyperparameters appear in the defined scenario: $a_\tau, b_\tau, \boldsymbol{\mu}_{\beta_g}, \Sigma_{\beta_g}, \boldsymbol{\mu}_{0g}, \Sigma_{0g}, \mathbb{B}_{\mathbb{D}_g}, d_{\mathbb{D}_g}, a_{\eta_g}, b_{\eta_g}, a_{\nu_g}, b_{\nu_g}, \boldsymbol{\mu}_{\alpha_g}, \Sigma_{\alpha_g}, \boldsymbol{a}_\pi, \boldsymbol{\mu}_{\boldsymbol{\xi}}$ and $\Sigma_{\boldsymbol{\xi}}$. For future derivation, all are considered as fixed and pre-specified parameters. Moreover, the choices of prior distributions for particular parameters are clearly depicted on Figure 3.1 with a help of DAG.

3.2.2 Posterior distributions

Using a Bayesian theorem, the joint posterior density of the latent variables and parameters of interest is of the form of:

$$\begin{aligned} p(\boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\theta} | \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}) &\propto p(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\theta}) \\ &= p(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta} | \boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\theta}) \underbrace{p(\boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\theta})}_{\text{prior}}. \end{aligned} \tag{3.12}$$

The prior distributions were specified in the previous section and prior independence is assumed. Under a hierarchical model, the joint density of data and parameters, i.e. a right side of (3.12), is defined as

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\theta}) &= \prod_{i=1}^{K} p(\boldsymbol{y}_i, t_i, \delta_i | V_i = g, \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i | \boldsymbol{\theta}) p(V_i | \boldsymbol{\theta}) p(\boldsymbol{\theta}) = \\ &= \prod_{i=1}^{K} p(\boldsymbol{y}_i | V_i = g, \boldsymbol{b}_i, \boldsymbol{\theta}_l) p(t_i, \delta_i | V_i = g, \boldsymbol{\theta}_s) p(\boldsymbol{b}_i | \boldsymbol{\theta}_l) \\ &\quad \mathrm{P}[V_i = g | \boldsymbol{\theta}_p] p(\boldsymbol{\theta}_s) p(\boldsymbol{\theta}_l) p(\boldsymbol{\theta}_p). \end{aligned} \tag{3.13}$$

The goal is to estimate the unknown parameters $\boldsymbol{\theta}$. The Bayesian estimator of the parameters is naturally a marginal posterior expected value $\mathbb{E}(\boldsymbol{\theta}_j | \boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\delta})$,

where $\boldsymbol{\theta}_j$ is a subvector of $\boldsymbol{\theta}$. To achieve this, of course, we have to calculate a marginal posterior distribution of $\boldsymbol{\theta}_j$. That requires calculation of the multidimensional integral $p(\boldsymbol{\theta}_j|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta}) = \int_{\boldsymbol{\Theta}_{-j}} p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta})d\boldsymbol{\theta}_{-j}$, where $\boldsymbol{\theta}_{-j}$ is a vector $\boldsymbol{\theta}$ without j-th component and $\boldsymbol{\Theta}_{-j}$ represents a support of $\boldsymbol{\theta}_{-j}$. Unfortunately that may not lead to a close form solution and the computation of the integral can be challenging. For instance, the general form of cumulative hazard in the likelihood function of our model is likely to prevent analytical derivation of the marginal posterior distribution for some parameters without the proposed additional assumptions. Even if it is possible to derive a marginal posterior distribution we cannot be sure that with a general specification of prior distributions it would not lead to some well-known distribution that would make the calculation of the posterior expectation easier.

Nevertheless, as it was already mentioned it is sufficient to compute full conditional distributions for the individual parameters of interest and then use a theory of MCMC and employ some algorithm such as the Gibbs algorithm or the Metropolis - Hastings algorithm to get results. The general form of the full conditional distribution is,

$$\begin{aligned} p(\boldsymbol{\theta}_j|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{b},\boldsymbol{V},\boldsymbol{\theta}_{-j}) &\propto p(\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta}|\boldsymbol{b},\boldsymbol{V},\boldsymbol{\theta})p(\boldsymbol{b}|\boldsymbol{V},\boldsymbol{\theta})p(\boldsymbol{V}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{V},\boldsymbol{\theta}_l)p(\boldsymbol{b}|\boldsymbol{V},\boldsymbol{\theta}_l)p(\boldsymbol{\theta}_l) \\ &\quad p(\boldsymbol{t},\boldsymbol{\delta}|\boldsymbol{V},\boldsymbol{\theta}_s)p(\boldsymbol{\theta}_s)p(\boldsymbol{V}|\boldsymbol{\theta}_p)p(\boldsymbol{\theta}_p), \end{aligned} \tag{3.14}$$

and we are going to present the full conditional distributions of all components of $\boldsymbol{\theta}$ and latent variables V and $\boldsymbol{b}$ in Chapter 4.

3.2.3 A selection of number of classes

The number of classes is another topic that needs to be discussed and there is not only one option of how to decide about the number of classes. While in the frequentist approach the use of Bayesian information criterion (BIC) is common and straightforward. It cannot be said that this criterion is also employed in Bayesian approach. Several different criteria are introduced and consequently compared to get the number of classes in Bayesian finite mixture models (Nasserinejad et al. [2017]). To our best knowledge there is no literature available where the comparison of criteria would be extended to Bayesian joint latent class models. Nevertheless the recommendation coming from the study are applied to the JLCM (Andrinopoulou et al. [2018]). The suggestion is to use Rosseau and Mengersen‘s criterion (Rousseau and Mengersen [2011]), this criterion is closely connected to the choice of the parameters in the Dirichlet prior distribution. Of the other criteria, let us mention the deviance information criterion (DIC) or the application of reversible jump MCMC algorithm (Nasserinejad et al. [2017]). However, we will not focus on solving this problem or finding the optimum criterion as this topic requires careful attention and explanations which is beyond the scope of this thesis.

4. Derivation of full conditional distributions

As mentioned in the previous chapter, the main motivation for deriving full conditional distributions is that it allows us to estimate the model parameters by using the Bayesian approach. To the best of our knowledge, it seems there is no specific software package for this type of model available except the general software for Bayesian modelling (e.g. `JAGS`). It means that the user is forced to derive it on his own if he wants to use the Bayesian method to estimate the parameters of the model. This chapter should provide the reader with the theoretical background that is needed for the implementation.

The chapter is divided into three sections corresponding to the parts of the model related to the different data type, i.e. the contribution of the longitudinal model, time-to-event model and class probability model,

$$p(\boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}) \propto p(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\theta}_l)p(\boldsymbol{t}, \boldsymbol{\delta}|\boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\theta}_s)p(\boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\theta}). \tag{4.1}$$

We start with the derivation of the full conditional distributions of the the parameters related to the model of the longitudinal outcome. We assume that the model introduced in Chapter 3 holds including the specification of prior distributions.

4.1 The longitudinal model

The contribution of the longitudinal model to posterior distribution of $\boldsymbol{\theta}_l$ and the latent variable $\boldsymbol{b}$ takes a form,

$$\begin{aligned}
&p(\boldsymbol{\theta}_l, \boldsymbol{b}|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}) \propto p(\boldsymbol{y}|\boldsymbol{V}, \boldsymbol{b}, \boldsymbol{\theta}_l)p(\boldsymbol{V}, \boldsymbol{b}, \boldsymbol{\theta}_l) \\
&\propto \prod_{g=1}^{G} \prod_{i:v_i=g} p(\boldsymbol{y}_i|V_i = g, \boldsymbol{b}_i, \boldsymbol{\theta}_l)p(\boldsymbol{b}_i|V_i = g, \boldsymbol{\theta}_l)p(\boldsymbol{\theta}_l) \\
&= \prod_{g=1}^{G} \prod_{i:v_i=g} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n_i}{2}} \exp\Big\{ -\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)^\mathsf{T}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)\Big\} \\
&\left(\frac{1}{2\pi}\right)^{\frac{q}{2}} |\mathbb{D}_g|^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2}(\boldsymbol{b}_i - \boldsymbol{\mu}_g)^\mathsf{T}\mathbb{D}_g^{-1}(\boldsymbol{b}_i - \boldsymbol{\mu}_g)\Big\} p(\tau)p(\boldsymbol{\beta}_g)p(\boldsymbol{\mu}_g)p(\mathbb{D}_g^{-1}),
\end{aligned} \tag{4.2}$$

because of the independence of the longitudinal model and the survival model given the group g. Only the priors of parameters τ, $\boldsymbol{\beta}_g$, $\boldsymbol{\mu}_g$ and $\mathbb{D}_g^{-1}$ appear in (4.2) due to the fact that other components of $\boldsymbol{\theta}$ are not involved in the longitudinal part of the model. In the following lines we derive a full conditional distribution of $\boldsymbol{\beta}_g$, τ, $\boldsymbol{b}_g$, $\boldsymbol{\mu}_g$ and $\mathbb{D}_g^{-1}$ given the class g, respectively.

Let us start with a vector of regression coefficients $\boldsymbol{\beta}_g$ describing the effect of particular covariates. The random vector $\boldsymbol{\beta}_g$ follows conditioning on the class g a prior distribution $\mathsf{N}_p\big(\boldsymbol{\mu}_{\beta g}, \Sigma_{\beta_g}\big)$. Denote $\boldsymbol{\theta}_{lg}$ as a set of parameters that are taking part in the longitudinal model for class g. We proceed from (4.2) and

because of the prior independence and independence of $\boldsymbol{\beta}_g$ and random effects $\boldsymbol{b}_i$ for $i = 1, \dots, K$, then we can write $p(\boldsymbol{\beta}_g|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots)$ as

$$
\begin{aligned}
p(\boldsymbol{\beta}_g|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) &\propto \prod_{i:v_i=g} p(\boldsymbol{y}_i|V_i = g, \boldsymbol{b}_i, \boldsymbol{\theta}_{lg}) p(\boldsymbol{\beta}_g) \\
&\propto \prod_{i:v_i=g} \left(\frac{1}{2\pi\sigma_e^2}\right)^{\frac{n_i}{2}} \exp\Big\{ -\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)^\mathsf{T}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)\Big\} \\
&\quad |\Sigma_{\boldsymbol{\beta}_g}|^{-1} \exp\Big\{ -\frac{1}{2}(\boldsymbol{\beta}_g - \boldsymbol{\mu}_{\boldsymbol{\beta}_g})^\mathsf{T}\Sigma_{\boldsymbol{\beta}_g}(\boldsymbol{\beta}_g - \boldsymbol{\mu}_{\boldsymbol{\beta}_g})\Big\} \\
&\propto \exp\Big\{ -\frac{1}{2}\Big[\underbrace{-2\sum_{i:v_i=g}(\mathbb{X}_i\boldsymbol{\beta}_g)^\mathsf{T}\Sigma^{-1}(\boldsymbol{y}_i - \mathbb{Z}_i\boldsymbol{b}_i) + \sum_{i:v_i=g}(\mathbb{X}_i\boldsymbol{\beta}_g)^\mathsf{T}\Sigma^{-1}(\mathbb{X}_i\boldsymbol{\beta}_g)}_{(\star)} \\
&\quad \underbrace{+\boldsymbol{\beta}_g^\mathsf{T}\Sigma_{\boldsymbol{\beta}_g}^{-1}\boldsymbol{\beta}_g - 2\boldsymbol{\beta}_g^\mathsf{T}\Sigma_{\boldsymbol{\beta}_g}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}_g}\Big]\Big\}}_{(\star)}.
\end{aligned}
$$

We omit the terms that do not depend on $\boldsymbol{\beta}_g$ and factor out $\boldsymbol{\beta}_g$ in $(\star)$,

$$
\begin{aligned}
(\star) = C - \frac{1}{2}\Big[\boldsymbol{\beta}_g^\mathsf{T} \underbrace{(\Sigma_{\boldsymbol{\beta}_g}^{-1} + \sum_{i:v_i=g}\mathbb{X}_i^\mathsf{T}\Sigma^{-1}\mathbb{X}_i)}_{\equiv\, \mathbb{A}_{\boldsymbol{\beta}_g}}\boldsymbol{\beta}_g \\
- 2\boldsymbol{\beta}_g^\mathsf{T} \underbrace{\Big(\sum_{i:v_i=g}\mathbb{X}_i^\mathsf{T}\Sigma^{-1}(\boldsymbol{y}_i - \mathbb{Z}_i\boldsymbol{b}_i) + \boldsymbol{\mu}_{\boldsymbol{\beta}_g}\Big)}_{\equiv\, \mathbf{B}_{\boldsymbol{\beta}_g}}\Big],
\end{aligned}
$$

where $\mathbb{A}_{\boldsymbol{\beta}_g}$ is a symmetric and positive-definite (PD) matrix when we assume that $\Sigma_{\boldsymbol{\beta}_g}$ is a symmetric and positive-definite matrix and it holds because it is a covariance matrix. The second term $\mathbb{X}_i^\mathsf{T}\Sigma^{-1}\mathbb{X}_i = \frac{1}{\sigma^2}\mathbb{X}_i^\mathsf{T}\mathbb{X}_i$ is a symmetric positive definite if $\mathbb{X}_i$ has a full column rank. Then $\mathbb{A}_{\boldsymbol{\beta}_g}$ is a sum of two symmetric positive definite matrices is again symmetric PD matrix. Then we can write $\mathbb{A}_{\boldsymbol{\beta}_g} = \mathbb{A}_{\boldsymbol{\beta}_g}^{\frac{1}{2}}\mathbb{A}_{\boldsymbol{\beta}_g}^{\frac{1}{2}}$. The term in brackets from $(\star)$ can be converted to the square, i.e.

$$
\begin{aligned}
(\star) &= \tilde{C} - \frac{1}{2}(\mathbb{A}_{\boldsymbol{\beta}_g}^{\frac{1}{2}}\boldsymbol{\beta}_g - \mathbb{A}_{\boldsymbol{\beta}_g}^{-\frac{1}{2}}\mathbf{B}_{\boldsymbol{\beta}_g}^\mathsf{T})^\mathsf{T}(\mathbb{A}_{\boldsymbol{\beta}_g}^{\frac{1}{2}}\boldsymbol{\beta}_g - \mathbb{A}_{\boldsymbol{\beta}_g}^{-\frac{1}{2}}\mathbf{B}_{\boldsymbol{\beta}_g}^\mathsf{T}) \\
&= \tilde{C} - \frac{1}{2}(\boldsymbol{\beta}_g - \mathbb{A}_{\boldsymbol{\beta}_g}^{-\frac{1}{2}}\mathbb{A}_{\boldsymbol{\beta}_g}^{-\frac{1}{2}}\mathbf{B}_{\boldsymbol{\beta}_g}^\mathsf{T})^\mathsf{T}\mathbb{A}_{\boldsymbol{\beta}_g}^{\frac{1}{2}}\mathbb{A}_{\boldsymbol{\beta}_g}^{\frac{1}{2}}(\boldsymbol{\beta}_g - \mathbb{A}_{\boldsymbol{\beta}_g}^{-\frac{1}{2}}\mathbb{A}_{\boldsymbol{\beta}_g}^{-\frac{1}{2}}\mathbf{B}_{\boldsymbol{\beta}_g}^\mathsf{T}) \\
&= \tilde{C} - \frac{1}{2}(\boldsymbol{\beta}_g - \mathbb{A}_{\boldsymbol{\beta}_g}^{-1}\mathbf{B}_{\boldsymbol{\beta}_g}^\mathsf{T})^\mathsf{T}\mathbb{A}_{\boldsymbol{\beta}_g}(\boldsymbol{\beta}_g - \mathbb{A}_{\boldsymbol{\beta}_g}^{-1}\mathbf{B}_{\boldsymbol{\beta}_g}^\mathsf{T})
\end{aligned}
$$

Together we have

$$
p(\boldsymbol{\beta}_g|\boldsymbol{y}, \boldsymbol{T}, \boldsymbol{\delta}, \dots) \propto \exp\Big\{ -\frac{1}{2}(\boldsymbol{\beta}_g - \mathbb{A}_{\boldsymbol{\beta}_g}^{-1}\mathbf{B}_{\boldsymbol{\beta}_g}^\mathsf{T})^\mathsf{T}\mathbb{A}_{\boldsymbol{\beta}_g}(\boldsymbol{\beta}_g - \mathbb{A}_{\boldsymbol{\beta}_g}^{-1}\mathbf{B}_{\boldsymbol{\beta}_g}^\mathsf{T})\Big\}. \tag{4.3}
$$

It follows from (4.3) that the full conditional distribution of $\boldsymbol{\beta}_g$ is $\mathsf{N}_p(\mathbb{A}_{\boldsymbol{\beta}_g}^{-1}\mathbf{B}_{\boldsymbol{\beta}_g}^\mathsf{T}, \mathbb{A}_{\boldsymbol{\beta}_g}^{-1})$.

The covariance matrix Σ is diagonal with σ^2 on the diagonal. So, we derive a full conditional distribution just for the inverse transformation of σ^2, i.e. $\tau = \sigma^{-2}$. Due to the fact that τ is independent of class membership we are allowed to derive

a full conditional distribution without a conditioning to a fixed class g. A prior distribution of τ is specified as $\Gamma(a_\tau, b_\tau)$ and is independent of other priors. Plug in this to (3.14) then the full conditional distribution of τ takes a form,

$$p(\tau|\mathbf{y}, t, \delta, \dots) \propto \prod_{g=1}^{G} \prod_{i:v_i=g} p(\boldsymbol{y}_i|V_i = g, \boldsymbol{b}_i, \boldsymbol{\theta}_l)p(\boldsymbol{\theta}_l)$$

where a_σ and b_σ are unknown parameters that we have to specified in advance. With particular densities, we obtain

$$\begin{aligned} p(\tau|\boldsymbol{y}, t, \delta, \dots) &\propto \prod_{g=1}^{G} \prod_{i:v_i=g} \tau^{a_\tau - 1} e^{-\tau b_\tau} \tau^{\frac{n_i}{2}} \\ &\quad \exp\Big\{ -\frac{\tau}{2}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)^\mathsf{T}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)\Big\} \\ &= \prod_{g=1}^{G} \tau^{(a_\tau + \frac{N_g}{2} - 1)} \exp\Big\{ -\tau\Big(b_\tau + \\ &\quad \sum_{i:v_i=g} \frac{1}{2}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)^\mathsf{T}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)\Big)\Big\}, \end{aligned}$$

the scaling constants are omitted and $\sum_{i:v_i=g} n_i = N_g$, the number of subject in the class N_g. Subsequently, $\sum_{g=1}^{G} \frac{N_g}{2} = \frac{N}{2}$ and, we write,

$$\begin{aligned} p(\tau|\boldsymbol{y}, t, \delta, \dots) &\propto \tau^{a_\tau + \sum_{g=1}^{G} \sum_{i:v_i=g} \frac{n_i}{2} - 1} \\ &\quad \exp\Big\{ -\tau \underbrace{\Big(b_\tau + \sum_{g=1}^{G} \sum_{i:v_i=g} \frac{1}{2}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)^\mathsf{T}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)\Big)}_{\equiv \tilde{b}_\tau} \Big\}, \end{aligned}$$

thus a full conditional distribution of τ while using all observed data is equal to $\Gamma\Big(a_\tau + \frac{N}{2}, \tilde{b}_\tau\Big)$, under the condition that both parameters of the distribution are positive. It holds because $\frac{N}{2}$ is positive as $N_g \in \mathbb{N}$ and the second parameter consists of sum of quadratic terms.

Next, we derive a full conditional distribution of a latent variable $\boldsymbol{b}$ for fixed g. This variable represents the random effects of individuals in the longitudinal model. For fixed g, where $\boldsymbol{b}_g$ is a vector, where the individual components are subvectors $\boldsymbol{b}_i$ such that $i \in \{j \in \mathbb{N} : V_j = g\}$. Again we proceed from (4.2). The prior distributions of the parameters that are independent of random effects are omitted,

$$p(\boldsymbol{b}_g|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \propto \prod_{i:v_i=g} p(\boldsymbol{y}_i|V_i = g, \boldsymbol{b}_i, \boldsymbol{\theta}_l)p(\boldsymbol{b}_i|V_i = g, \boldsymbol{\theta}_l).$$

Plug in the corresponding densities and omit scaled constants.

$$
\begin{aligned}
p(\boldsymbol{b}_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) &\propto \prod_{i:v_i=g} \exp\Big\{ -\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)^\mathsf{T}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)\Big\} \\
&\quad \exp\Big\{ -\frac{1}{2}(\boldsymbol{b}_i - \boldsymbol{\mu}_g)^\mathsf{T}\mathbb{D}_g^{-1}(\boldsymbol{b}_i - \boldsymbol{\mu}_g)\Big\} \\
&\propto \prod_{i:v_i=g} \exp\Big\{ -\frac{1}{2}\Big(\underbrace{\boldsymbol{b}_i^\mathsf{T}\mathbb{Z}_i^\mathsf{T}\Sigma^{-1}\mathbb{Z}_i\boldsymbol{b}_i - 2\boldsymbol{b}_i^\mathsf{T}\Sigma^{-1}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g)}_{(\star)} \\
&\quad \underbrace{+\boldsymbol{b}_i^\mathsf{T}\mathbb{D}_g^{-1}\boldsymbol{b}_i - 2\boldsymbol{b}_i^\mathsf{T}\mathbb{D}_g^{-1}\boldsymbol{\mu}_g}_{(\star)}\Big)\Big\}
\end{aligned}
$$

Follow similar steps as we did when we derived the full conditional distribution of $\boldsymbol{\beta}$, factor out $\boldsymbol{b}_i$

$$
(\star) = \; C + \boldsymbol{b}_i^\mathsf{T}\underbrace{(\mathbb{Z}_i^\mathsf{T}\Sigma^{-1}\mathbb{Z}_i + \mathbb{D}_g^{-1})}_{\equiv\ \mathbb{E}_{ig}}\boldsymbol{b}_i - 2\boldsymbol{b}_i^\mathsf{T}\Big(\underbrace{\Sigma^{-1}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g) + \mathbb{D}_g^{-1}\boldsymbol{\mu}_g}_{\equiv\ \mathbf{F}_{ig}}\Big),
$$

where $\mathbb{Z}_i^\mathsf{T}\Sigma^{-1}\mathbb{Z}_i$ is a symmetric and positive-definite matrix when we assume that Σ is a symmetric and positive-definite matrix. It holds because it is a covariance matrix. Multiplication by a design matrix $\mathbb{Z}_i$ does not change anything if $\mathbb{Z}_i$ has a full column rank. The matrix $\mathbb{D}_g$ is a covariance matrix, thus it is a symmetric and positive definite from the definition. Then $\mathbb{E}_{ig}$ is a sum of two symmetric positive definite matrices and is also a symmetric positive-definite matrix. Then we can write $\mathbb{E}_{ig} = \mathbb{E}_{ig}^{\frac{1}{2}}\mathbb{E}_{ig}^{\frac{1}{2}}$ and we can rewrite $(\star)$ such that,

$$
\begin{aligned}
(\star) &= \; \tilde{C} + \Big(\mathbb{E}_{ig}^{\frac{1}{2}}\boldsymbol{b}_i - \mathbb{E}_{ig}^{-\frac{1}{2}}\mathbf{F}_{ig}^\mathsf{T}\Big)^\mathsf{T}\Big(\mathbb{E}_{ig}^{\frac{1}{2}}\boldsymbol{b}_i - \mathbb{E}_{ig}^{-\frac{1}{2}}\mathbf{F}_{ig}^\mathsf{T}\Big) \\
&= \; \tilde{C} + \Big(\boldsymbol{b}_i - \mathbb{E}_{ig}^{-\frac{1}{2}}\mathbb{E}_{ig}^{-\frac{1}{2}}\mathbf{F}_{ig}^\mathsf{T}\Big)^\mathsf{T}\mathbb{E}_{ig}^{\frac{1}{2}}\mathbb{E}_{ig}^{\frac{1}{2}}\Big(\boldsymbol{b}_i - \mathbb{E}_{ig}^{-\frac{1}{2}}\mathbb{E}_{ig}^{-\frac{1}{2}}\mathbf{F}_{ig}^\mathsf{T}\Big) \\
&= \; \tilde{C} + \Big(\boldsymbol{b}_i - \mathbb{E}_{ig}^{-1}\mathbf{F}_{ig}^\mathsf{T}\Big)^\mathsf{T}\mathbb{E}_{ig}\Big(\boldsymbol{b}_i - \mathbb{E}_{ig}^{-1}\mathbf{F}_{ig}^\mathsf{T}\Big).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
p(\boldsymbol{b}_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) &\propto \prod_{i:v_i=g} \exp\Big\{ -\frac{1}{2}\Big(\boldsymbol{b}_i - \mathbb{E}_{ig}^{-1}\mathbf{F}_{ig}^\mathsf{T}\Big)^\mathsf{T}\mathbb{E}_{ig}\Big(\boldsymbol{b}_i - \mathbb{E}_{ig}^{-1}\mathbf{F}_{ig}^\mathsf{T}\Big)\Big\} \\
&\propto \exp\Big\{ -\frac{1}{2}\sum_{i:v_i=g}\Big(\boldsymbol{b}_i - \mathbb{E}_{ig}^{-1}\mathbf{F}_{ig}^\mathsf{T}\Big)^\mathsf{T}\mathbb{E}_{ig}\Big(\boldsymbol{b}_i - \mathbb{E}_{ig}^{-1}\mathbf{F}_{ig}^\mathsf{T}\Big)\Big\} \\
&\propto \exp\Big\{ -\frac{1}{2}\Big(\boldsymbol{b}_g - \mathbb{E}_g^{-1}\mathbf{F}_g^\mathsf{T}\Big)^\mathsf{T}\mathbb{E}_g\Big(\boldsymbol{b}_g - \mathbb{E}_g^{-1}\mathbf{F}_g^\mathsf{T}\Big)\Big\},
\end{aligned}
$$

where $\mathbb{E}_g$ is a block matrix consisted of submatrices $\mathbb{E}_{ig}$, such that $i \in \{j \in \mathbb{N} : V_j = g\}$ on the diagonal, zeros otherwise, as a consequence of mutual independence of $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$ for $i \neq j$. A vector $\mathbf{F}_g$ is composed out of subvectors $\mathbf{F}_{ig}, , i \in \{j \in \mathbb{N} : V_j = g\}$. To sum up, a full conditional distribution of $\boldsymbol{b}_g$ is $\mathsf{N}_{qk_g}(\mathbb{E}_g^{-1}\mathbf{F}_g^\mathsf{T}, \mathbb{E}_g^{-1})$, where $k_g = \sum_{i=1}^K \mathbb{1}(V_i = g)$ and a full conditional distribution of $\boldsymbol{b}_i$ is $\mathsf{N}_q(\mathbb{E}_{ig}^{-1}\mathbf{F}_{ig}^\mathsf{T}, \mathbb{E}_{ig}^{-1})$.

There are two more parameters of the longitudinal model performing like the parameters of the distribution of the random effects, i.e. a vector of expected values $\boldsymbol{\mu}_g$ and the covariance matrix $\mathbb{D}_g$ given a class g. First, we derive a full conditional distribution for $\boldsymbol{\mu}_g$. A prior of $\boldsymbol{\mu}_g$ is $\mathsf{N}_q(\boldsymbol{\mu}_{0g}, \Sigma_{0g})$, so then we have,

$$\begin{aligned}
p(\boldsymbol{\mu}_g|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) &\propto \prod_{i:v_i=g} p(\boldsymbol{b}_i|V_i = g, \mathbb{D}_g, \boldsymbol{\mu}_g) p(\boldsymbol{\mu}_g) \\
&\propto \prod_{i:v_i=g} \exp\Big\{ -\frac{1}{2}(\boldsymbol{b}_i - \boldsymbol{\mu}_g)^\mathsf{T} \mathbb{D}_g^{-1} (\boldsymbol{b}_i - \boldsymbol{\mu}_g) \Big\} \\
&\quad \exp\Big\{ -\frac{1}{2}(\boldsymbol{\mu}_g - \boldsymbol{\mu}_{0g})^\mathsf{T} \Sigma_{0g}^{-1} (\boldsymbol{\mu}_g - -\boldsymbol{\mu}_{0g}) \Big\} \\
&\propto \exp\Big\{ -\frac{1}{2}\Big(\underbrace{-2 \sum_{i:v_i=g} \boldsymbol{\mu}_g^\mathsf{T} \mathbb{D}_g^{-1} \boldsymbol{b}_i + k_g \boldsymbol{\mu}_g^\mathsf{T} \mathbb{D}_g^{-1} \boldsymbol{\mu}_g}_{(\star)} \\
&\quad \underbrace{-2\boldsymbol{\mu}_g^\mathsf{T} \Sigma_{0g}^{-1} \boldsymbol{\mu}_{0g} + \boldsymbol{\mu}_g^\mathsf{T} \Sigma_{0g}^{-1} \boldsymbol{\mu}_g)}_{(\star)} \Big) \Big\},
\end{aligned}$$

where $\sum_{i:v_i=g} 1 = \sum_{i=1}^{K} \mathbb{1}(V_i = g) = k_g$. Next, we apply the same steps as we already did several times, thus, we factor out $\boldsymbol{\mu}_g$.

$$(\star) = \; C - 2\boldsymbol{\mu}_g^\mathsf{T} \underbrace{(\Sigma_{0g}^{-1} + \sum_{i:v_i=g} \mathbb{D}_g^{-1} \boldsymbol{b}_i)}_{\equiv \mathbf{J}_g} + \boldsymbol{\mu}_g^\mathsf{T} \underbrace{(\Sigma_{0g}^{-1} + k_g \mathbb{D}_g^{-1})}_{\equiv \mathbb{K}_g} \boldsymbol{\mu}_g,$$

where $\mathbb{K}_g$ is a symmetric positive definite matrix due to the fact that it is a sum of two symmetric positive matrices as Σ_{0g} and $\mathbb{D}_g$ are covariance matrices. Thus, $\mathbb{K}_g = \mathbb{K}_g^{\frac{1}{2}} \mathbb{K}_g^{\frac{1}{2}}$ and by converting $(\star)$ into the square, we get

$$\begin{aligned}
(\star) &= \tilde{C} + (\mathbb{K}_g^{\frac{1}{2}} \boldsymbol{\mu}_g - \mathbb{K}_g^{-\frac{1}{2}} \mathbf{J}_g^\mathsf{T})^\mathsf{T} (\mathbb{K}_g^{\frac{1}{2}} \boldsymbol{\mu}_g - \mathbb{J}_g^{-\frac{1}{2}} \mathbf{J}_g^\mathsf{T}) \\
&= \tilde{C} + (\boldsymbol{\mu}_g - \mathbb{K}_g^{-1} \mathbf{J}_g^\mathsf{T})^\mathsf{T} \mathbb{K}_g (\boldsymbol{\mu}_g - \mathbb{K}_g^{1} \mathbf{J}_g^\mathsf{T}).
\end{aligned}$$

To sum up,

$$p(\boldsymbol{\mu}_g|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \propto \exp\Big\{ -\frac{1}{2}(\boldsymbol{\mu}_g - \mathbb{K}_g^{-1} \mathbf{J}_g^\mathsf{T})^\mathsf{T} \mathbb{K}_g (\boldsymbol{\mu}_g - \mathbb{K}_g^{1} \mathbf{J}_g^\mathsf{T}) \Big\},$$

a full conditional distribution of $\boldsymbol{\mu}_g$ is again q-dimensional normal distribution, i.e., $\mathsf{N}_q(\mathbb{K}_g^{-1} \mathbf{J}_g^\mathsf{T}, \mathbb{K}_g^{-1})$.

Last but not least we need to derive a full conditional derivation for a covariance matrix $\mathbb{D}_g$. However, we will not derive it for the covariance matrix, but for the inverse of the covariance matrix. A prior distribution of inverse of the covariance matrix is $\mathsf{W}_q(d_{\mathbb{D}_g}, \mathbb{B}_{\mathbb{D}_g})$. By using the same process as before, we obtain,

$$
\begin{aligned}
p(\mathbb{D}_g^{-1}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) &\propto \prod_{i:v_i=g} p(\boldsymbol{b}_i|V_i=g,\mathbb{D}_g^{-1},\boldsymbol{\mu}_g)p(\mathbb{D}_g^{-1}) \\
&\propto \prod_{i:v_i=g} |\mathbb{D}_g|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)^\mathsf{T}\mathbb{D}_g^{-1}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)\Big\} \\
&\quad |\mathbb{D}_g^{-1}|^{\frac{d_{\mathbb{D}_g}-q-1}{2}} \exp\Big\{-\frac{1}{2}\mathrm{tr}(\mathbb{B}_{\mathbb{D}_g}^{-1}\mathbb{D}_g^{-1})\Big\} \\
&\propto |\mathbb{D}_g^{-1}|^{\frac{d_{\mathbb{D}_g}-q}{2}} \exp\Big\{-\frac{1}{2}\underbrace{\Big(\sum_{i:v_i=g}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)^\mathsf{T}\mathbb{D}_g^{-1}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)+\mathrm{tr}(\mathbb{B}_{\mathbb{D}_g}^{-1}\mathbb{D}_g^{-1})\Big)}_{(\star)}\Big\}.
\end{aligned}
$$

Next, we rewrite a term $(\star)$ using a properties of a trace of a matrix (APPENDIX A.2), thus,

$$
\begin{aligned}
(\star) &= \mathrm{tr}\Big(\sum_{i:v_i=g}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)^\mathsf{T}\mathbb{D}_g^{-1}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)\Big)+\mathrm{tr}(\mathbb{B}_{\mathbb{D}_g}^{-1}\mathbb{D}_g^{-1}) \\
&= \sum_{i:v_i=g}\mathrm{tr}\Big((\boldsymbol{b}_i-\boldsymbol{\mu}_g)(\boldsymbol{b}_i-\boldsymbol{\mu}_g)^\mathsf{T}\mathbb{D}_g^{-1}\Big)+\mathrm{tr}(\mathbb{B}_{\mathbb{D}_g}^{-1}\mathbb{D}_g^{-1}) \\
&= \mathrm{tr}\Big(\sum_{i:v_i=g}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)(\boldsymbol{b}_i-\boldsymbol{\mu}_g)^\mathsf{T}\mathbb{D}_g^{-1}+\mathbb{B}_{\mathbb{D}_g}^{-1}\mathbb{D}_g^{-1}\Big) \\
&= \mathrm{tr}\Big\{\underbrace{\Big(\sum_{i:v_i=g}(\boldsymbol{b}_i-\boldsymbol{\mu}_g)(\boldsymbol{b}_i-\boldsymbol{\mu}_g)^\mathsf{T}+\mathbb{B}_{\mathbb{D}_g}^{-1}\Big)}_{\tilde{\mathbb{B}}_{\mathbb{D}_g}^{-1}}\mathbb{D}_g^{-1}\Big\}.
\end{aligned}
$$

By putting together the previous results, we get

$$
p(\mathbb{D}_g^{-1}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto |\mathbb{D}_g^{-1}|^{\frac{d_{\mathbb{D}_g}-q}{2}} \exp\Big\{-\frac{1}{2}\mathrm{tr}\Big(\tilde{\mathbb{B}}_{\mathbb{D}_g}^{-1}\mathbb{D}_g^{-1}\Big)\Big\}, \tag{4.4}
$$

so that it is again a Wishart distribution. To conclude, for given class g, the density$p(\mathbb{D}_g^{-1}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots)$ is proportional to $\mathsf{W}_q(d_{\mathbb{D}_g}+1,\tilde{\mathbb{B}}_{\mathbb{D}_g})$ if $\tilde{\mathbb{B}}_{\mathbb{D}_g}$ is positive definite. It holds because $\mathbb{B}_{\mathbb{D}_g}$ is a positive definite matrix and the second term is a sum of quadratic terms that are nonnegative. Moreover, $d_{\mathbb{D}_g}+1$ should be greater then q and $d_{\mathbb{D}_g}>q$ from definitions, thus, this condition is satisfied.

In the case of the longitudinal model we derived a full conditional distribution for $\boldsymbol{\beta}_g,\tau,\boldsymbol{b}_i,\boldsymbol{\mu}_g$ and $\mathbb{D}_g^{-1}$, for all of them we figured out that the form of derived densities was proportional to densities of some standard distributions. That makes the estimation of the model easier and it allows us to use a Gibbs algorithm because it is possible to generate easily from those standard distributions. The results are summarized in the following lemma.

Lemma 1. *Lets assume that the model defined in section 3.2 holds, then the full conditional distributions of the parameters related to the longitudinal model are*

$$
p(\boldsymbol{\beta}_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \sim \mathsf{N}_p(\mathbb{A}_{\beta_g}^{-1}\mathbf{B}_{\beta_g}^\mathsf{T},\mathbb{A}_{\beta_g}^{-1}),
$$

where $\mathbb{A}_{\beta_g}=\Sigma_{\boldsymbol{\beta}_g}^{-1}+\sum_{i:v_i=g}\mathbb{X}_i^\mathsf{T}\Sigma^{-1}\mathbb{X}_i$ *and* $\mathbf{B}_{\beta_g}=\sum_{i:v_i=g}\mathbb{X}_i^\mathsf{T}\Sigma^{-1}(\boldsymbol{y}_i-\mathbb{Z}_i\boldsymbol{b}_i)+\boldsymbol{\mu}_{\beta_g}$,

$$p(\tau|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \sim \Gamma\Big(a_\tau + \frac{N}{2}, \tilde{b}_\tau\Big),$$

where $\tilde{b}_\tau = b_\tau + \frac{1}{2}\sum_{g=1}^{G}\sum_{i:v_i=g}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)^\mathsf{T}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g - \mathbb{Z}_i\boldsymbol{b}_i)$,

$$p(\boldsymbol{b}_i|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \sim \mathsf{N}_q(\mathbb{E}_{ig}^{-1}\mathbf{F}_{ig}^\mathsf{T}, \mathbb{E}_{ig}^{-1}),$$

where $\mathbb{E}_{ig} = \mathbb{Z}_i^\mathsf{T}\Sigma^{-1}\mathbb{Z}_i + \mathbb{D}_g^{-1}$ *and* $\mathbf{F}_{ig} = \Sigma^{-1}(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\beta}_g) + \mathbb{D}_g^{-1}\boldsymbol{\mu}_g$,

$$p(\boldsymbol{\mu}_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \sim \mathsf{N}_q(\mathbb{K}_g^{-1}\mathbf{J}_g^\mathsf{T}, \mathbb{K}_g^{-1})$$

where $\mathbb{K}_g = (\Sigma_{0g}^{-1} + k_g\mathbb{D}_g^{-1})$ *and* $\mathbf{J}_g = \Sigma_{0g}^{-1} + \sum_{i:v_i=g}\mathbb{D}_g^{-1}\boldsymbol{b}_i$ *and as the last,*

$$p(\mathbb{D}_g^{-1}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \sim \mathsf{W}_q(d_{\mathbb{D}_g} + 1, \tilde{\mathbb{B}}_{\mathbb{D}_g}),$$

where $\tilde{\mathbb{B}}_{\mathbb{D}_g} = \sum_{i:v_i=g}(\boldsymbol{b}_i - \boldsymbol{\mu}_g)(\boldsymbol{b}_i - \boldsymbol{\mu}_g)^\mathsf{T} + \mathbb{B}_{\mathbb{D}_g}^{-1}$.

Proof. See previous derivations. □

4.2 The survival model

In this part we focus on the full conditional distributions for parameters of the survival model. By $\boldsymbol{\theta}_{sg}$ we denote a set of parameters for the survival model that are class-specific. Note that not necessarily all of the parameters depend on the class however it is allowed to be like that in contrast to a standard deviation σ from the longitudinal model which never depends on the class.

We proceed from (4.1). By the definition of a hierarchical model and conditional independence, the part of (4.1) corresponding to the contribution of the longitudinal model and class probabilities is constant with respect to the parameters of interest, thus it is not needed in the following derivations. We assume that λ_{0g} follows a priori Weibull distribution. Then the contribution of the survival model to the posterior distribution of $\boldsymbol{\theta}_s$ is of the form

$$\begin{aligned}
&p(\boldsymbol{\theta}_s|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta}) \propto p(\boldsymbol{t},\boldsymbol{\delta}|\boldsymbol{V},\boldsymbol{\theta}_s)p(\boldsymbol{V},\boldsymbol{\theta}_s)\\
&\prod_{i:v_i=g} p(t_i,\delta_i|V_i=g,\boldsymbol{\theta}_s)p(\boldsymbol{\theta}_s) =\\
&\prod_{i:v_i=g} \lambda_i(t_i|V_i=g,\boldsymbol{\theta}_s)^{\delta_i}S(t_i|V_i=g,\boldsymbol{\theta}_{sg})p(\boldsymbol{\theta}_{sg}) =\\
&\prod_{i:v_i=g} \Big\{\lambda_{0g}(t_i,\nu_g,\eta_g)\exp(\tilde{\mathbf{X}}_i(t)^\mathsf{T}\boldsymbol{\alpha}_g)\Big\}^{\delta_i}\exp\Big(-\Lambda_i(t_i|\nu_g,\eta_g,\boldsymbol{\alpha}_g,\tilde{\mathbf{X}}_i(t)\Big) \qquad (4.5)\\
&p(\nu_g)p(\eta_g)p(\boldsymbol{\alpha}_g) = \prod_{i:v_i=g}\Big\{\nu_g\eta_g t_i^{\nu_g-1}\exp(\tilde{\mathbf{X}}_i(t)^\mathsf{T}\boldsymbol{\alpha}_g)\Big\}^{\delta_i}\\
&\exp\Big(-\nu_g\eta_g\int_0^{t_i} s^{\nu_g-1}e^{\tilde{\mathbf{X}}_i(s)^\mathsf{T}\boldsymbol{\alpha}_g}ds\Big)p(\nu_g)p(\eta_g)p(\boldsymbol{\alpha}_g).
\end{aligned}$$

On the next lines the full conditional distributions $p(\eta_g|\boldsymbol{y},\dots)$, $p(\nu_g|\boldsymbol{y},\dots)$, and $p(\boldsymbol{\alpha}_g|\boldsymbol{y},\dots)$ are derived. The full conditional distributions of parameters of η_g and ν_g depend on the parametrization of λ_{0g}, due to the fact that those parameters are parameters of the Weibull distribution, i.e. with different parametrization we have to derive a full conditional distribution for corresponding parameters

corresponding to the actual parametrization. However, the full conditional distribution of $\boldsymbol{\alpha}_g$ is independent of the parametrization of a baseline hazard λ_{0g}, thus it can be used even if the baseline hazard is a piecewise constant or the B-splines are employed.

First, we derive a full conditional distribution for a scale parameter η_g. The prior distribution $p(\eta_g)$ was defined as $\Gamma(a_{\eta g}, b_{\eta g})$. The prior independence of η_g, ν_g and $\boldsymbol{\alpha}_g$ is assumed and all other parameters are considered as a constant, so then it follows from (4.5),

$$p(\eta_g|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \propto \prod_{i:v_i=g} \eta_g^{\delta_i} \exp\Big\{ -\eta_g \nu_g \int_0^{t_i} s^{\nu_g - 1} \exp(\tilde{\mathbf{X}}_i(s)^\mathsf{T} \boldsymbol{\alpha}_g) ds \Big\} \eta_g^{a_{\eta_g} - 1} e^{-b_{\eta_g} \eta_g}.$$

Denote $\nu_g \int_0^{t_i} s^{\nu_g - 1} \exp(\tilde{\mathbf{X}}_i(s)^\mathsf{T} \boldsymbol{\alpha}_g) ds = C_{ig}$, and merge the exponents together, then

$$p(\eta_g|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \propto \eta_g^{\sum_{i:v_i=g} \delta_i + a_{\eta_g} - 1} \exp\Big\{ -\eta_g (b_{\eta_g} + \sum_{i:v_i=g} C_{ig}) \Big\}.$$

This is a well-known density of Gamma distribution without a normalized constant. Thus, a full conditional distribution of η_g follows $\Gamma\big(\sum_{i:v_i=g} \delta_i + a_{\eta g}, b_{\eta g} + \sum_{i:v_i=g} C_{ig}\big)$ if both parameters of Gamma distribution are positive. The term $\sum_{i:v_i=g} \delta_i + a_{\eta g}$ is positive because $\sum_{i:v_i=g} \delta_i \geq 0$ is a sum of indicators and $a_{\eta g}$ was defined as a positive parameter. The second term is also positive because $C_{ig} = \frac{\Lambda_i(T_i|\boldsymbol{\theta}_{sg})}{\eta_g}$ where $\Lambda_i(T_i|\boldsymbol{\theta}_{sg})$ is nonnegative from the definition of the cumulative hazard function, $\eta_g > 0$ and $b_{ig} > 0$ because they are parameters of the Gamma distribution or Weibull distribution.

The second parameter of a baseline hazard is a shape parameter ν_g of the Weibull distribution. The full conditional distribution of ν_g for fixed class g with a prior $p(\nu_g) \sim \Gamma(a_{\nu_g}, b_{\nu_g})$ plugged in (4.5) is of the form,

$$p(\nu_g|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \propto \prod_{i:v_i=g} (\nu_g t_i^{\nu_g - 1})^{\delta_i} \exp\Big\{ -\eta_g \nu_g \int_0^{t_i} s^{\nu_g - 1} \exp(\tilde{\mathbf{X}}_i(s)^\mathsf{T} \boldsymbol{\alpha}_g) ds \Big\} \nu_g^{a_{\nu_g} - 1} e^{-b_{\nu_g} \nu_g}.$$

Next, we reorder the terms to try to factor out ν_g,

$$\begin{aligned} p(\nu_g|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) &\propto \prod_{i:v_i=g} \nu_g^{\delta_i + a_{\nu_g} - 1} \exp\Big\{ \delta_i (\nu_g - 1) \log t_i \\ &\quad - \nu_g \Big(\eta_g \int_0^{t_i} s^{\nu_g - 1} \exp(\tilde{\mathbf{X}}_i(s)^\mathsf{T} \boldsymbol{\alpha}_g) ds + b_{\nu_g} \Big) \Big\} \\ &\propto \nu_g^{\sum_{i:v_i=g} \delta_i + a_{\nu_g} - 1} \exp\Big\{ \nu_g \Big(\sum_{i:v_i=g} \delta_i \log t_i \\ &\quad - \eta_g \sum_{i:v_i=g} \int_0^{t_i} s^{\nu_g - 1} \exp(\tilde{\mathbf{X}}_i(s)^\mathsf{T} \boldsymbol{\alpha}_g) ds - b_{\nu_g} \Big) \Big\}. \end{aligned}$$

Unfortunately, the full conditional distribution of ν_g does not seem to be from the family of well-known distributions. Even if it is assumed that $\tilde{\mathbf{X}}_i(s) = \tilde{\mathbf{X}}_i$

does not depend on time and it is possible to compute an integral, the term is not simplified enough to reach a form of some known distribution, i.e we have

$$p(\nu_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto \nu_g^{\sum_{i:v_i=g}\delta_i+a_{\nu_g}-1}\exp\Big\{\nu_g\Big(\sum_{i:v_i=g}\delta_i\log t_i - b_{\nu_g}\Big) - \eta_g\sum_{i:v_i=g}\exp(\tilde{\mathbf{X}}_i^{\mathsf{T}}\boldsymbol{\alpha}_g)t_i^{\nu_g}\Big\}.$$

The results lead to an application of the a different algorithm than the Gibbs algorithm, e.g. the Metropolis-Hastings algorithm, when we want to estimate the parameters, because we are not able to generate easily from this distribution.

The last group of parameters is a vector of αs that explains the effect of covariates to the occurrence of event. Again we proceed from (4.5) and all terms that does not depend on $\boldsymbol{\alpha}_g$ are omitted and $p(\boldsymbol{\alpha}_g)$ is plugged in. The prior distribution is defined as $\boldsymbol{\alpha}_g \sim \mathsf{N}_r(\boldsymbol{\mu}_{\boldsymbol{\alpha}_g}, \Sigma_{\boldsymbol{\alpha}_g})$, thus we obtain,

$$p(\boldsymbol{\alpha}_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto \prod_{i:v_i=g}\Big\{\exp(\tilde{\mathbf{X}}_i(t)^{\mathsf{T}}\boldsymbol{\alpha}_g)\Big\}^{\delta_i} \exp\Big(-\nu_g\eta_g\int_0^{t_i} s^{\nu_g-1}e^{\tilde{\mathbf{X}}_i(s)^{\mathsf{T}}\boldsymbol{\alpha}_g}ds\Big)\exp\Big(-\frac{1}{2}(\boldsymbol{\alpha}_g-\boldsymbol{\mu}_{\boldsymbol{\alpha}_g})^{\mathsf{T}}\Sigma_{\boldsymbol{\alpha}_g}^{-1}(\boldsymbol{\alpha}_g-\boldsymbol{\mu}_{\boldsymbol{\alpha}_g})\Big),$$

To simplify the term, we denote by $C_{i\alpha_g} = -\nu_g\eta_g\sum_{i:v_i=g}\int_0^{t_i} s^{\nu_g-1}e^{\tilde{\mathbf{X}}_i(s)^{\mathsf{T}}\boldsymbol{\alpha}_g}ds$ and put together all remaining terms in exponent where $\boldsymbol{\alpha}$ is present, i.e.

$$p(\boldsymbol{\alpha}_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto \exp\Big(C_{i\alpha_g} + \underbrace{\sum_{i:v_i=g}\delta_i\tilde{\mathbf{X}}_i(t)^{\mathsf{T}}\boldsymbol{\alpha}_g - \frac{1}{2}(\boldsymbol{\alpha}_g^{\mathsf{T}}\Sigma_{\boldsymbol{\alpha}_g}^{-1}\boldsymbol{\alpha}_g - 2\boldsymbol{\mu}_{\boldsymbol{\alpha}_g}^{\mathsf{T}}\Sigma_{\boldsymbol{\alpha}_g}^{-1}\boldsymbol{\alpha}_g)}_{(\star)}\Big),$$

as a consequence of presence of $C_{i\alpha_g}$, there is no possibility to adjust the exponent into the form where $\boldsymbol{\alpha}_g$ can be fully factored out. At least $(\star)$ can be converted into the square,

$$\begin{aligned}
(\star) &= \sum_{i:v_i=g}\delta_i\tilde{\mathbf{X}}_i(t)^{\mathsf{T}}\boldsymbol{\alpha}_g - \frac{1}{2}(\boldsymbol{\alpha}_g^{\mathsf{T}}\Sigma_{\boldsymbol{\alpha}_g}^{-1}\boldsymbol{\alpha}_g - 2\boldsymbol{\mu}_{\boldsymbol{\alpha}_g}^{\mathsf{T}}\Sigma_{\boldsymbol{\alpha}_g}^{-1}\boldsymbol{\alpha}_g)\\
&= -\frac{1}{2}\Big(\boldsymbol{\alpha}_g^{\mathsf{T}}\Sigma_{\boldsymbol{\alpha}_g}^{-1}\boldsymbol{\alpha}_g - 2\underbrace{(\boldsymbol{\mu}_{\boldsymbol{\alpha}_g}^{\mathsf{T}}\Sigma_{\boldsymbol{\alpha}_g}^{-1} + \sum_{i:v_i=g}\delta_i\tilde{\mathbf{X}}_i(t)^{\mathsf{T}})}_{\equiv\ \mathbf{D}_{\boldsymbol{\alpha}_g}}\boldsymbol{\alpha}_g\Big)\\
&= -\frac{1}{2}\Big(\Sigma_{\boldsymbol{\alpha}_g}^{-\frac{1}{2}}\boldsymbol{\alpha}_g - \Sigma_{\boldsymbol{\alpha}_g}^{\frac{1}{2}}\mathbf{D}_{\boldsymbol{\alpha}_g}^{\mathsf{T}}\Big)^{\mathsf{T}}\Big(\Sigma_{\boldsymbol{\alpha}_g}^{-\frac{1}{2}}\boldsymbol{\alpha}_g - \Sigma_{\boldsymbol{\alpha}_g}^{\frac{1}{2}}\mathbf{D}_{\boldsymbol{\alpha}_g}^{\mathsf{T}}\Big) + C\\
&= -\frac{1}{2}\Big(\boldsymbol{\alpha}_g - \Sigma_{\boldsymbol{\alpha}_g}\mathbf{D}_{\boldsymbol{\alpha}_g}^{\mathsf{T}}\Big)^{\mathsf{T}}\Sigma_{\boldsymbol{\alpha}_g}^{-1}\Big(\boldsymbol{\alpha}_g - \Sigma_{\boldsymbol{\alpha}_g}\mathbf{D}_{\boldsymbol{\alpha}_g}^{\mathsf{T}}\Big) + C,
\end{aligned}$$

where $\Sigma_{\boldsymbol{\alpha}_g}^{\frac{1}{2}}\Sigma_{\boldsymbol{\alpha}_g}^{\frac{1}{2}} = \Sigma_{\boldsymbol{\alpha}_g}$ and C is a constant with respect to the full conditional distribution, i.e. there is no $\boldsymbol{\alpha}_g$ involved. All together we obtain,

$$p(\boldsymbol{\alpha}_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto \exp\Big\{-\frac{1}{2}\Big(\boldsymbol{\alpha}_g - \Sigma_{\boldsymbol{\alpha}_g}\mathbf{D}_{\boldsymbol{\alpha}_g}^{\mathsf{T}}\Big)^{\mathsf{T}}\Sigma_{\boldsymbol{\alpha}_g}^{-1}\Big(\boldsymbol{\alpha}_g - \Sigma_{\boldsymbol{\alpha}_g}\mathbf{D}_{\boldsymbol{\alpha}_g}^{\mathsf{T}}\Big) - C_{i\alpha_g}\Big\}.$$

Nevertheless, we run into the same problem as in the case of ν_g because of the cumulative hazard function, especially as it is not possible to rearrange a term $C_{i\alpha_g}$ in a way that would allow us to find any simple form of full conditional distribution. Note that if this term is omitted, the expression corresponds to a normal distribution.

To conclude, we did not obtain any form of standard distribution for $p(\nu_g|\boldsymbol{y},\dots)$ and $p(\boldsymbol{\alpha}_g|\boldsymbol{y},\dots)$, on the other hand, $p(\eta_g|\boldsymbol{y},\dots)$ belongs to the family of standard distributions. As a consequence it is not possible to employ a Gibbs algorithm due to the fact that we are not able to easily generate from the full conditional distributions. Instead, we propose the usage of the Metropolis-Hasting algorithm as an alternative. The results of this section are summarized in the following lemma.

Lemma 2. *Lets assume that the model defined in section 3.2 holds, then the full conditional distributions of the parameters related to the survival model are*

$$p(\eta_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \sim \Gamma\Big(\sum_{i:v_i=g}\delta_i + a_{\eta g}, b_{\eta g} + \sum_{i:v_i=g} C_{ig}\Big),$$

where $C_{ig} = \nu_g \int_0^{t_i} s^{\nu_g-1} exp(\tilde{\mathbf{X}}_i(s)^\mathsf{T}\boldsymbol{\alpha}_g)ds$,

$$p(\nu_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto \nu_g^{\sum_{i:v_i=g}\delta_i + a_{\nu_g}-1} exp\Big\{\nu_g\Big(\sum_{i:v_i=g}\delta_i log t_i - b_{\nu_g}\Big) - \eta_g \sum_{i:v_i=g} exp(\tilde{\mathbf{X}}_i^\mathsf{T}\boldsymbol{\alpha}_g)t_i^{\nu_g}\Big\},$$

and as the last,

$$p(\boldsymbol{\alpha}_g|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto exp\Big\{-\frac{1}{2}\Big(\boldsymbol{\alpha}_g - \Sigma_{\boldsymbol{\alpha}_g}\mathbf{D}_{\boldsymbol{\alpha}_g}^\mathsf{T}\Big)^\mathsf{T}\Sigma_{\boldsymbol{\alpha}_g}^{-1}\Big(\boldsymbol{\alpha}_g - \Sigma_{\boldsymbol{\alpha}_g}\mathbf{D}_{\boldsymbol{\alpha}_g}^\mathsf{T}\Big) - C_{i\alpha_g}\Big\},$$

where $C_{i\alpha_g} = -\nu_g\eta_g \sum_{i:v_i=g}\int_0^{t_i} s^{\nu_g-1} e^{\tilde{\mathbf{X}}_i(s)^\mathsf{T}\boldsymbol{\alpha}_g}ds$, $\mathbf{D}_{\boldsymbol{\alpha}_g} = \boldsymbol{\mu}_{\boldsymbol{\alpha}_g}^\mathsf{T}\Sigma_{\boldsymbol{\alpha}_g}^{-1} + \sum_{i:v_i=g}\delta_i\tilde{\mathbf{X}}_i(t)^\mathsf{T}$.

Proof. See previous derivations. □

4.3 The class probability model

The last part of our model is associated with class membership. Class probability can be modeled in several ways. First, we derive a full conditional distribution of the latent variable V_i that is assumed to follow a discrete distribution $\mathcal{A}(\boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i = (\pi_{1i},\dots,\pi_{1G})^\mathsf{T}$. Then we focus on the full conditional distribution of parameters that occurs in definition of class probability, if those are modelled as functions of covariates using (2.7).

The joint full conditional distribution for $\boldsymbol{V}$ is proportional to,

$$p(\boldsymbol{V}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto \prod_{g=1}^{G}\prod_{i=1}^{K}\Big\{p(\boldsymbol{y}_i|\boldsymbol{b}_i,\boldsymbol{\theta}_l)p(\boldsymbol{b}_i|\boldsymbol{\theta}_l)p(t_i,\delta_i|\boldsymbol{\theta}_s)\Big\}^{\mathbb{1}(V_i=g)} \mathrm{P}[V_i=g|\boldsymbol{\theta}_p]^{\mathbb{1}(V_i=g)}p(\boldsymbol{\theta}_l)p(\boldsymbol{\theta}_s)p(\boldsymbol{\theta}_p),$$

because of the prior independence, the priors of $\boldsymbol{\theta}$ can be omitted. Subsequently, put together all terms with an indicator in the exponent,

$$p(\boldsymbol{V}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto \prod_{g=1}^{G}\prod_{i=1}^{K}\Big\{\underbrace{p(\boldsymbol{y}_i|\boldsymbol{b}_i,\boldsymbol{\theta}_g)p(\boldsymbol{b}_i|\boldsymbol{\theta}_g)p(t_i,\delta_i|\boldsymbol{\theta}_s)\mathrm{P}[V_i=g|\boldsymbol{\theta}_p]}_{\propto\tilde{\pi}_{ig}}\Big\}^{\mathbb{1}(V_i=g)}.$$

The expression above represents a density of the joint full conditional distribution of $\boldsymbol{V}$ except a normalized constant. The random variables $V_i,\ i=1,\dots,K$ are independent. The probabilities $\tilde{\pi}_{ig}$ are proportional to the product of densities $p(\boldsymbol{y}_i|\boldsymbol{b}_i,\boldsymbol{\theta}_{lg})p(\boldsymbol{b}_i|\boldsymbol{\theta}_{lg})p(t_i,\delta_i|\boldsymbol{\theta}_{sg})$. This product is equal to $\tilde{\pi}_{ig}$ after the standardization. Afterwards, it is possible to claim that the full conditional distribution of V_i is again a discrete distribution, multinomial distribution, with probabilities $\tilde{\boldsymbol{\pi}}_i=(\tilde{\pi}_{i1},\dots,\tilde{\pi}_{iG})^{\mathsf{T}}$, such that $\tilde{\pi}_{ig}\in(0,1)$ for all $i=1,\dots,K$ and $g=1,\dots G$, and $\sum_{g=1}^{G}\tilde{\pi}_{ig}=1$.

4.3.1 Modeling class probabilities

By the end of this chapter, we will discuss probability modeling. In the first case we consider that probabilities $\boldsymbol{\pi}$ do not depend on other parameters, in the second case we model probabilities through a multinomial logistic regression as (3.3). In both cases we differentiate between two options, first, the probabilities do not depend on the subject, i.e. $\boldsymbol{\pi}_i=\boldsymbol{\pi}$ for $i=1,\dots,K$, or the probabilities are subject specific, i.e. $\boldsymbol{\pi}_i\neq\boldsymbol{\pi}_j$, for $i\neq j$.

For the case where the class-membership probability is not subject specific $\pi_{ig}=\pi_g$ and as a prior we choose a Dirichlet distribution, i.e.

$$\begin{aligned}
p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K}\pi_g^{\mathbb{1}(V_i=g)}p(\boldsymbol{\pi}) = \\
&\prod_{i=1}^{K}\pi_g^{\mathbb{1}(V_i=g)}\frac{1}{B(\boldsymbol{a}_\pi)}\prod_{g=1}^{G}\pi_g^{a_{\pi_g}-1} = \\
&\pi_g^{\sum_{i=1}^{K}\mathbb{1}(V_i=g)}\frac{1}{B(\boldsymbol{a}_\pi)}\prod_{g=1}^{G}\pi_g^{a_{\pi_g}-1} = \\
&\frac{1}{B(\boldsymbol{a}_\pi)}\prod_{g=1}^{G}\pi_g^{\sum_{i=1}^{K}\mathbb{1}(V_i=g)+a_{\pi_g}-1}.
\end{aligned}$$

It follows that a full conditional distribution of $\boldsymbol{\pi}$ is again a Dirichlet distribution, however, with parameters $\tilde{\boldsymbol{a}}_\pi=(\sum_{i=1}^{K}\mathbb{1}(V_i=1)+a_{\pi_1},\dots,\sum_{i=1}^{K}\mathbb{1}(V_i=G)+a_{\pi_G})^{\mathsf{T}}$.

Remark. If the probabilities are subject specific, by the same steps as above we derive a full conditional distribution for $\boldsymbol{\pi}_i$, such that

$$\begin{aligned}
p(\boldsymbol{\pi}_i|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) &\propto \pi_g^{\mathbb{1}(V_i=g)}p(\boldsymbol{\pi}_i) = \pi_{ig}^{\mathbb{1}(V_i=g)}\frac{1}{B(\boldsymbol{a}_{\pi_i})}\prod_{g=1}^{G}\pi_{ig}^{a_{\pi_{ig}}-1} = \\
&\frac{1}{B(\boldsymbol{a}_{\pi_i})}\prod_{g=1}^{G}\pi_{ig}^{\mathbb{1}(V_i=g)+a_{\pi_{ig}}-1}.
\end{aligned}$$

Then, the full conditional distribution of $\boldsymbol{\pi}_i$ is a Dirichlet distribution with parameters $\tilde{\boldsymbol{a}}_{\pi_i}=(\mathbb{1}(V_i=1)+a_{\pi_{i1}},\dots,\mathbb{1}(V_i=G)+a_{\pi_{iG}})^{\mathsf{T}}$.

Now assume that probabilities take a form of (3.3). First, assume that probabilities are not subject specific, i.e. do not depend on i. This can be understood as a special case of multinomial logistic regression with only one G-level covariate that classifies the individuals to the classes. Later, we generalize this approach to the dependency of the set of covariates.

The expression (3.3) implies that a parameter $\boldsymbol{\pi}$ depends on parameters $\boldsymbol{\xi}$ and a prior of $\boldsymbol{\xi}$ is a multivariate normal for the reason that $\boldsymbol{\xi}$ performs as a vector of effects of the covariates. Thus, the full conditional distribution of $\boldsymbol{\xi}$ takes the following form,

$$\begin{aligned} p(\boldsymbol{\pi}(\boldsymbol{\xi})|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) &\propto \prod_{g=1}^{G}\prod_{i=1}^{K}\pi_g(\boldsymbol{\xi})^{\mathbb{1}(V_i=g)}p(\boldsymbol{\pi}(\boldsymbol{\xi})) \\ &\propto \prod_{g=1}^{G}\prod_{i=1}^{K}\Big(\frac{e^{\xi_g}}{\sum_{l=1}^{G}e^{\xi_l}}\Big)^{\mathbb{1}(V_i=g)}|\Sigma_\xi|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\mu}_\xi)^{\mathsf{T}}\Sigma_\xi^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu}_\xi)\Big) \end{aligned}$$

The denominator in a probability term that is independent of i and it is the same for all g, thus, it can be factored out and we rearrange the prior of $\boldsymbol{\xi}$,

$$\begin{aligned} p(\boldsymbol{\xi}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto &\Big(\frac{1}{\sum_{l=1}^{G}e^{\xi_l}}\Big)^{K}\prod_{g=1}^{G}\sigma_{\xi_g}^{-1}\exp\underbrace{\Big(\xi_g\sum_{i=1}^{K}\mathbb{1}(V_i=g)\Big)}_{(\star)} \\ &\underbrace{\exp\Big(-\frac{1}{2}\frac{(\xi_g-\mu_{\xi_g})^2}{\sigma_{\xi_g}^2}\Big)}_{(\star)}. \end{aligned}$$

Subsequently, the term $(\star)$ is converted into the square for each g fixed separately, i.e.,

$$\begin{aligned} (\star) &= \xi_g\sum_{i=1}^{K}\mathbb{1}(V_i=g)-\frac{1}{2}\sum_{g=1}^{G}\frac{(\xi_g-\mu_{\xi_g})^2}{\sigma_{\xi_g}^2} \\ &= \xi_g\sum_{i=1}^{K}\mathbb{1}(V_i=g)-\frac{1}{2}\Big(\frac{\xi_g^2}{\sigma_{\xi_g}^2}-2\frac{\xi_g\mu_{\xi_g}}{\sigma_{\xi_g}^2}\Big)+C \\ &= -\frac{1}{2\sigma_{\xi_g}^2}\Big(\xi_g^2-2\xi_g(\mu_{\xi_g}+\sigma_{\xi_g}^2\sum_{i=1}^{K}\mathbb{1}(V_i=g)\Big)+C \\ &= -\frac{\Big(\xi_g-(\mu_{\xi_g}+\sigma_{\xi_g}^2\sum_{i=1}^{K}\mathbb{1}(V_i=g))\Big)^2}{2\sigma_{\xi_g}^2}+\tilde{C}. \end{aligned}$$

By substituting $(\star)$ and omitting $\tilde{C}$, which is independent of ξ_g, we get the full conditional distribution of $\boldsymbol{\xi}$ except for the normalized constant,

$$\begin{aligned} p(\boldsymbol{\xi}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta},\dots) \propto &\Big(\frac{1}{\sum_{l=1}^{G}e^{\xi_l}}\Big)^{K}\prod_{g=1}^{G}\sigma_{\xi_g}^{-1} \\ &\exp\Big\{-\frac{1}{2}\frac{\Big(\xi_g-(\mu_{\xi_g}+\sigma_{\xi_g}^2\sum_{i=1}^{K}\mathbb{1}(V_i=g))\Big)^2}{\sigma_{\xi_g}^2}\Big\} \end{aligned}$$

Sadly, as a result of the presence of the first fraction, the derived density does not belong to the family of standard distributions.

When $\boldsymbol{\pi}$ is a subject specific, the situation is analogous to the one just derived. Nevertheless ξ_g is not scalar but a vector of a dimension m and there is a set of covariates that are included in the probability term as it was defined in (3.3).

Now, $\boldsymbol{\xi}_g = (\xi_{0g}, \dots, \xi_{m-1,G})^\mathsf{T}$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\mathsf{T}, \dots, \boldsymbol{\xi}_G^\mathsf{T})^\mathsf{T}$ and $p(\boldsymbol{\xi}) = p(\boldsymbol{\xi}_1) \cdot \dots \cdot p(\boldsymbol{\xi}_G)$ because $p(\boldsymbol{\xi}) \sim \mathsf{N}_{mG}(\boldsymbol{\mu}_{\boldsymbol{\xi}}, \Sigma_{\boldsymbol{\xi}})$ where $\Sigma_{\boldsymbol{\xi}}$ is a block-matrix with $m \times m$ matrices $\Sigma_{\boldsymbol{\xi}_g}$ on diagonal and zeros otherwise. $\Sigma_{\boldsymbol{\xi}_g} = \mathrm{diag}(\sigma_{\xi_{1g}^2}, \dots, \sigma^2_{\xi_{mg}})$, it follows that $\Sigma_{\boldsymbol{\xi}}$ is a diagonal matrix and the individual components of $\boldsymbol{\xi}$ are independent with respect to a prior distribution. It follows that a full conditional distribution takes a form,

$$
\begin{aligned}
p(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) &\propto \prod_{i=1}^{K}\prod_{g=1}^{G} \pi_{ig}(\boldsymbol{\xi})^{\mathbb{1}(V_i=g)} p(\boldsymbol{\xi}) \\
&\propto \prod_{i=1}^{K}\prod_{g=1}^{G} \Big(\frac{e^{\boldsymbol{\xi}_g^\mathsf{T}\tilde{\mathbf{X}}_i}}{\sum_{l=1}^{G} e^{\boldsymbol{\xi}_l^\mathsf{T}\tilde{\mathbf{X}}_i}}\Big)^{\mathbb{1}(V_i=g)} |\Sigma_{\boldsymbol{\xi}}|^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\mu}_{\boldsymbol{\xi}})^\mathsf{T}\Sigma_{\boldsymbol{\xi}}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu}_{\boldsymbol{\xi}})\Big) \\
&\propto \frac{1}{\prod_{i=1}^{K}\sum_{l=1}^{G} e^{\boldsymbol{\xi}_l^\mathsf{T}\tilde{\mathbf{X}}_i}} \prod_{g=1}^{G} \exp\Big(\underbrace{\sum_{i:v_i=g} \boldsymbol{\xi}_g^\mathsf{T}\tilde{\mathbf{X}}_i - \frac{1}{2}(\boldsymbol{\xi}_g-\boldsymbol{\mu}_{\boldsymbol{\xi}_g})^\mathsf{T}\Sigma_{\boldsymbol{\xi}_g}^{-1}(\boldsymbol{\xi}_g-\boldsymbol{\mu}_{\boldsymbol{\xi}_g})}_{(\star)} \Big).
\end{aligned}
\tag{4.6}
$$

For fixed g we put the terms in the exponent together and analogously to the previous case we get

$$
\begin{aligned}
(\star) &= \sum_{i:v_i=g} \boldsymbol{\xi}_g^\mathsf{T}\tilde{\mathbf{X}}_i - \frac{1}{2}(\boldsymbol{\xi}_g^\mathsf{T}\Sigma_{\boldsymbol{\xi}_g}^{-1}\boldsymbol{\xi}_g - 2\boldsymbol{\xi}_g^\mathsf{T}\Sigma_{\boldsymbol{\xi}_g}^{-1}\boldsymbol{\mu}_{\boldsymbol{\xi}_g}) + C \\
&= -\frac{1}{2}[\boldsymbol{\xi}_g^\mathsf{T}\Sigma_{\boldsymbol{\xi}_g}^{-1}\boldsymbol{\xi}_g - 2\boldsymbol{\xi}_g^\mathsf{T}(\Sigma_{\boldsymbol{\xi}_g}^{-1}\boldsymbol{\mu}_{\boldsymbol{\xi}_g} + \sum_{i:v_i=g}\tilde{\mathbf{X}}_i)] + C \\
&= -\frac{1}{2}\Big[\Big(\boldsymbol{\xi}_g - (\boldsymbol{\mu}_{\boldsymbol{\xi}_g} + \Sigma_{\boldsymbol{\xi}_g}^{-1}\sum_{i:v_i=g}\tilde{\mathbf{X}}_i)\Big)^\mathsf{T}\Sigma_{\boldsymbol{\xi}_g}\Big(\boldsymbol{\xi}_g - \underbrace{(\boldsymbol{\mu}_{\boldsymbol{\xi}_g} + \Sigma_{\boldsymbol{\xi}_g}\sum_{i:v_i=g}\tilde{\mathbf{X}}_i)}_{\tilde{\boldsymbol{\mu}}_{\boldsymbol{\xi}_g}}\Big)\Big] + \tilde{C},
\end{aligned}
$$

where C and $\tilde{C}$ are constants with respect to $\boldsymbol{\xi}_g$. So if $(\star)$ is plugged to (4.6), $\tilde{C}$ is omitted, then the full conditional distribution of $\boldsymbol{\xi}$ is as follows,

$$
\begin{aligned}
&p(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \propto \\
&\qquad \frac{1}{\prod_{i=1}^{K}\sum_{l=1}^{G} e^{\boldsymbol{\xi}_l^\mathsf{T}\tilde{\mathbf{X}}_i}} \prod_{g=1}^{G} \exp\Big(-\frac{1}{2}[(\boldsymbol{\xi}_g-\tilde{\boldsymbol{\mu}}_{\boldsymbol{\xi}_g})^\mathsf{T}\Sigma_{\boldsymbol{\xi}_g}^{-1}(\boldsymbol{\xi}_g-\tilde{\boldsymbol{\mu}}_{\boldsymbol{\xi}_g})]\Big)
\end{aligned}
$$

Unfortunately neither for subject specific probabilities nor for probabilities independent of subjects we are able to classify the density as a one from the family of the standard distributions. We must resort to the Metropolis-Hastings algorithm or some other type of algorithm where there is no need to generate from full conditional distribution.

To conclude, the full conditional distributions for all Bayesian parameters where derived in this chapter, we also provided the reader with recommendations

of which type of algorithm should be use to estimate the model. In the next chapter these theoretical results are used to calculate a practical example. The results of this section are summarized in the following lemma.

Lemma 3. *Lets assume that the model defined in section 3.2 holds, then the full conditional distributions of the parameters related to the probability class model are*

$$p(V_i|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \sim \mathsf{Mult}_G(\tilde{\boldsymbol{\pi}}_i),$$

where $\tilde{\boldsymbol{\pi}}_i = (\tilde{\pi}_{i1}, \dots, \tilde{\pi}_{iG})^\mathsf{T}$ *for* $i = 1, \dots, K$ *and,* $\tilde{\pi}_{ig} \propto p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}_g)p(\boldsymbol{b}_i|\boldsymbol{\theta}_g)$ $p(t_i, \delta_i|\boldsymbol{\theta}_s)P[V_i = g|\boldsymbol{\theta}_p]$,

$$p(\boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \sim \mathsf{Dir}(\tilde{\boldsymbol{a}}_\pi),$$

where $\tilde{\boldsymbol{a}}_\pi = (\sum_{i=1}^K \mathbb{1}(V_i = 1) + a_{\pi_1}, \dots, \sum_{i=1}^K \mathbb{1}(V_i = G) + a_{\pi_G})^\mathsf{T}$, *if the probabilities are subject-specific then,*

$$p(\boldsymbol{\pi}_i|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \sim \mathsf{Dir}(\tilde{\boldsymbol{a}}_{\pi_i}),$$

where $\tilde{\boldsymbol{a}}_{\pi_i} = (\mathbb{1}(V_i = 1) + a_{\pi_{i1}}, \dots, \mathbb{1}(V_i = G) + a_{\pi_{iG}})^\mathsf{T}$, *and the last but not least,*

$$p(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\delta}, \dots) \propto \frac{1}{\prod_{i=1}^K \sum_{l=1}^G e^{\boldsymbol{\xi}_l^\mathsf{T} \tilde{\mathbf{X}}_i}} \prod_{g=1}^G exp\Big(-\frac{1}{2}[(\boldsymbol{\xi}_g - \tilde{\boldsymbol{\mu}}_{\boldsymbol{\xi}_g})^\mathsf{T} \Sigma_{\boldsymbol{\xi}_g}^{-1} (\boldsymbol{\xi}_g - \tilde{\boldsymbol{\mu}}_{\boldsymbol{\xi}_g})]\Big),$$

where $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\xi}_g} = \boldsymbol{\mu}_{\boldsymbol{\xi}_g} + \Sigma_{\boldsymbol{\xi}_g} \sum_{i:v_i=g} \tilde{\mathbf{X}}_i$.

Proof. See previous derivations. □

5. A simulation study

In the last chapter, we would like to provide the reader with a short simulation study in which we would like to focus on the ability of the model to distinguish between classes and then compare the ability of this approach to estimate parameters depending on the sample size. We use `JAGS` (Plummer [2003]) and a statistical software R (R Core Team [2017]) to compute the results of the simulation study. First we describe the data simulation process and then present the results of the study.

5.1 A description of the situation

We now describe the data simulation process and provide the model that was used to simulate the data. This model is based on a joint latent class model. We simulate 4 different situations. For each situation we have 100 data sets, the situations differ in the number of observations. We assumed three classes in our population, i.e. $G = 3$. The proportions of the classes in the simulated data sets and the number of subjects for the individual scenarios are summarized in Table 5.1. In order to better imagine the situation, we define a model that corresponds to real data situation - a clinical trial on patients with schizophrenia.

First, we start with the description of the longitudinal part of the model. We observe K subjects, patients with a schizophrenia, for 12 weeks, i.e. 84 days. During this period, each patient should complete a form 10 times about their mental fitness and the symptoms of the disease, a so-called PANSS (Positive and Negative Syndrome Scale). This longitudinal marker leads to the response vector $\mathbf{Y}$. These 10 individual time points were simulated evenly on the period of 84 days. Approximately half of the subjects were treated with a drug A and the rest were treated with a new drug B. The effect of the new drug B differs between classes and we also allow a different time effect in all three classes, the intercept and random effects are common to all classes. Thus, the model matrix $\mathbb{X}$ for fixed effects has three columns: intercept, drug (`drugB`) and time (`t`) and the model matrix $\mathbb{Z}$ for random effect has two columns: intercept and time (`t`). The general version of the model is,

$$\text{Class } g: \quad \texttt{PANSS}_i = \beta_0 + \beta_{1g}\texttt{drugB}_i + \beta_{2g}\texttt{t}_i + b_{0i} + b_{1i}\texttt{t}_i + \boldsymbol{\varepsilon}_i,$$

Table 5.1: Summary of the number of subjects in different classes per scenario, π is a proportion of the class with respect to the whole population.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Class 1 ($\pi_1 = 0.2$)	20	40	80	120
Class 2 ($\pi_2 = 0.5$)	50	100	200	300
Class 3 ($\pi_3 = 0.3$)	30	60	120	180
K^*	100	200	400	600

*Total number of subjects

The models we used for the simulation are the following,

$$\begin{aligned}
\text{Class 1:} \quad & \texttt{PANSS}_i = 120 + 0.02\texttt{drugB}_i - 0.2\texttt{t}_i + b_{0i} + b_{1i}\texttt{t}_i + \boldsymbol{\varepsilon}_i, \\
\text{Class 2:} \quad & \texttt{PANSS}_i = 120 + 5\texttt{drugB}_i + 0.5\texttt{t}_i + b_{0i} + b_{1i}\texttt{t}_i + \boldsymbol{\varepsilon}_i, \\
\text{Class 3:} \quad & \texttt{PANSS}_i = 120 - 10\texttt{drugB}_i - 0.5\texttt{t}_i + b_{0i} + b_{1i}\texttt{t}_i + \boldsymbol{\varepsilon}_i,
\end{aligned}$$

where $\boldsymbol{b}_i \sim \mathsf{N}_2(\mathbf{0}, \mathbb{D}), \mathbb{D} = diag(0.2, 0.25)$ and $\varepsilon_{ij} \sim \mathsf{N}(0, 6)$. There is $n_k = 10$ observation for each subject, however this number was reduced for some subjects because of censoring. The mechanism of generating censoring times is specified in the following paragraph.

The survival part of the model is represented by the Cox model with one factor variable. We include an effect of drug B, similarly to the longitudinal model. The effect of this variable differs between the classes. The baseline hazard follows a Weibull distribution and parameters are not class specific. As an event, we can imagine the occurrence of serious mental problems with an urgent need to see a doctor. The general version of the model is,

$$\text{Class } g: \quad \lambda_i(\texttt{T}_i^*) = \lambda_{i0}(\alpha_g \texttt{T}_i^*, \eta, \nu)\exp(\texttt{drugB}_i),$$

The models that were used to simulate an event time are defined as,

$$\begin{aligned}
\text{Class 1:} \quad & \lambda_i(\texttt{T}_i^*) = \lambda_{i0}(\texttt{T}_i^*, \eta, \nu)\exp(0.4\texttt{drugB}_i), \\
\text{Class 2:} \quad & \lambda_i(\texttt{T}_i^*) = \lambda_{i0}(\texttt{T}_i^*, \eta, \nu)\exp(-0.18\texttt{drugB}_i), \\
\text{Class 3:} \quad & \lambda_i(\texttt{T}_i^*) = \lambda_{i0}(\texttt{T}_i^*, \eta, \nu)\exp(-0.59\texttt{drugB}_i),
\end{aligned}$$

where $\nu = 1.1$ and $\eta = 1/e^5$ are the parameters of Weibull distribution (Appendix A.1).

Next, we generated censoring times $\texttt{C}_i$ for each subject as a $\min(\mathsf{U}(60, 90), 84))$, where $\mathsf{U}(60, 90)$ represents Uniform distribution on the interval $(60, 90)$. Then we took $\texttt{T}_i = \min(\texttt{T}_i^*, \texttt{C}_i)$ and $\delta_i = \mathbb{1}(\texttt{T}_i \leq \texttt{C}_i)$. The final modification of the data was that we omitted observations that were observed after censoring time per each subject from the data set generated in the first part.

As a statistical tool to analyze the data we define a model using `JAGS` and then we compute the estimates of the parameters of interest. We specified the prior distribution as weakly informative. The possible choices of weak informative prior distributions were discussed in Subsection 3.2.1, i.e. for the regression coefficients and the expected values of random effects we opted for a normal distribution with zero mean and inverse variance equal to 0.001. Then the gamma distribution was used as a prior for parameter of Weibull distribution, inverse variance of error terms and Wishart distribution was employed as a prior of the inverse covariance matrix of random effects. For class probability we use a Dirichlet distribution with parameters equal to $1/G$.

As a `JAGS` setting for computation of Markov chains, we select 1000 steps of the simulation as an adaptation to the model, then the length of the burn-in was $B = 5000$ and finally, the length of the Markov chain that was used to compute the estimates was chosen as $M = 25000$. However, we decided to use the thinning interval in order to try to decrease an autocorrelation and we set up $\texttt{thin} = 2$. In total, three Markov chains were generated for each data set.

Table 5.2: MC means of posterior means of non-class-specific parameters for all scenarios. The MC based standard error (SE) is in brackets on the line below the means.

	$\hat{\sigma}$	$\hat{\beta}$	$\hat{\eta}$	$\hat{\nu}$	$\hat{d}_{11}$	$\hat{d}_{21}$	$\hat{d}_{22}$
Scenario 1							
Estimate	6.05	118.64	2742.3	4.05	0.03	0.00	0.03
SD	(0.17)	(1.55)	(22678)	(0.78)	(0.01)	(0.01)	(0.01)
Scenario 2							
Estimate	6.06	118.63	25.69	4.51	0.03	0.00	0.03
SD	(0.11)	(0.88)	(171.34)	(0.57)	(0.00)	(0.01)	(0.01)
Scenario 3							
Estimate	6.07	118.82	0.62	4.85	0.03	0.00	0.03
SD	(0.07)	(1.04)	(4.29)	(0.36)	(0.01)	(0.00)	(0.01)
Scenario 4							
Estimate	6.06	118.50	0.00	5.15	0.03	0.00	0.03
SD	(0.06)	(1.69)	(0.00)	(0.30)	(0.01)	(0.01)	(0.01)

5.2 Results

In this section we summarized the results of the simulation study. As it was already mentioned, three Markov chains for each data set were generated and the estimates of the parameters were computed from these values. Unfortunately, `JAGS` was not able to finish a computing procedure in all cases. Some values were generated close to zero, it probably happens when the algorithm classifies the subject to the wrong class and then the density at the generated values is close to zero which creates problems in the algorithm and it stops. Using `JAGS` we are not able to prevent these situations. Thus there is always mentioned a final number of data set that we used to obtain presented estimates.

The estimates are summarized in Table 5.2, presented are the parameters that are not class specific, i.e regression coefficient from the longitudinal model (β_0), standard deviation of the error terms (σ), components of the covariance matrix $\mathbb{D}$, and parameters of the Weibull distribution (η, ν), and the class specific parameters are displayed in Table 5.3, i.e., the regression coefficients (β_{1g}, β_{2g} and α_g) and class probabilities (π_1, π_2 and π_3).

For the first scenario, we have on average $N = 885$ observation per $K = 100$ subjects and in approximately half of the cases per data set an event occurs. The algorithm was working for all 100 data sets. In the second scenario, there is on average $N = 1794$ observations for $K = 200$ subjects. Unfortunately, we obtained the estimates just for 92 data sets. Next, we have $N = 3622$ observations on average when $K = 400$ subjects in the data set and the algorithm was able to finish the calculations for 98 data sets. The data set in the fourth scenario has on average $N = 5425$ observations for $K = 600$ subjects and the algorithm converged to the solution for 99 samples.

Now we would like to compare the behavior of the estimates of the parameters. First, we evaluate the model ability to differentiate between the classes. The true

Table 5.3: MC means of posterior means of class-specific parameters for all scenarios. The MC based standard error (SE) is in brackets.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\pi}$
Scenario 1($K = 100$)				
Class 1	0.74 (1.31)	0.17 (0.47)	−7.81 (5.25)	0.31 (0.05)
Class 2	1.37 (1.61)	0.25 (0.51)	−7.85 (5.20)	0.44 (0.05)
Class 3	−6.25 (2.04)	−0.38 (0.46)	−2.09 (5.58)	0.27 (0.07)
Scenario 2 ($K = 200$)				
Class 1	0.79 (1.31)	0.10 (0.34)	−5.02 (4.29)	0.30 (0.04)
Class 2	0.61 (1.49)	0.36 (0.36)	−4.89 (4.12)	0.44 (0.04)
Class 3	−6.26 (2.06)	−0.41 (0.37)	−2.95 (4.54)	0.26 (0.06)
Scenario 3 ($K = 400$)				
Class 1	0.61 (1.08)	0.10 (0.26)	−1.64 (2.24)	0.31 (0.03)
Class 2	0.17 (1.28)	0.23 (0.24)	−1.43 (2.11)	0.43 (0.04)
Class 3	−6.65 (1.65)	−0.39 (0.24)	−0.75 (1.82)	0.26 (0.04)
Scenario 4 ($K = 600$)				
Class 1	−0.05 (1.49)	0.10 (0.23)	−0.54 (1.17)	0.31 (0.03)
Class 2	0.25 (1.34)	0.24 (0.21)	−0.45 (0.96)	0.42 (0.02)
Class 3	−6.64 (1.80)	−0.37 (0.21)	−0.21 (0.53)	0.27 (0.04)

value of the probability vector $\boldsymbol{\pi}$ from Table 5.1 is compared to the last column of Table 5.3. The model is able to recognize the largest class (class 2), however it underestimates the proportion of this class in the population in all scenarios. We got the estimates between 0.42 and 0.44 compared to the true value of 0.5. Looking at the estimated regression coefficients in Table 5.3. Furthermore, the model probably had a problem to differentiate between class 1 and 2 and some of the subjects that belong originally to the class 2 were wrongly classified as a class 1. This can be the reason why $\hat{\pi}_1$ overestimated the true value (0.31 compared to 0.2). Also the subjects from class 3 were most likely sometimes classified to the wrong class ($\hat{\pi}_3 = 0.26$ or 0.27 and $\pi_3 = 0.3$), as a class 1. Nevertheless the classification to class 3 was the most precise in contrast to the others.

The intercept β is close to the true value ($\hat{\beta} = 118.7$ in contrast to $\beta = 120$) and also the estimates of variance of the error terms are around 6 that is the true value of σ. For these parameters the model works well even if the number of subjects is the smallest possible (Scenario 1, $K = 100$). The components of covariance matrix $\mathbb{D}$ are underestimated in all scenarios. Moving on to the class-specific parameters in the longitudinal model, it is visible that all of the estimates in class 2 and 3 are shrinked towards zero in contrast to the true values of the parameters. As it was already mentioned class 1 is probably the most infected by the subject the from different classes, most likely by subjects from class 2, thus the estimates are influenced by this fact and they are larger than the true values. As a consequence of the misclassification of some subjects, the class-specific estimates approach the average across the classes of the true parameter values according to how strong the misclassification was. Due to the opposite signs of the true values of the coefficients, the estimates seem to be shrinked towards zero.

To conclude, the increasing number of subjects does not have such a huge impact on the longitudinal part of the model (regression coefficients and components of covariance matrix of random effects) and estimation class probabilities. We observed just the lower variance, however the estimates themselves do not differ so much.

Completely different behavior is noticed in the parameters of the Cox model. It is probably linked up with the unstable behavior of an estimate of η. Mainly, for small sample sizes we obtained values on the scale from close to zero to tens thousands. With an increasing sample size is $\hat{\eta}$ shrinked towards zero. The estimates of ν are quite stable however far away from the true value of the parameter. The estimates of α_g are strongly underestimated for small sample size. Similarly, as $\hat{\eta}$, $\hat{\alpha}_g$ are shrinked towards zero for larger sample sizes. The difference is in sign, while $\hat{\eta}$ is positive, $\hat{\alpha}_g$ are negative. This behavior can be rising from the definition of the model. The parameter η plays the role of intercept $\alpha_0 = log(\eta)$ in $\exp(\alpha_0 + \alpha_g \texttt{drugB}_i)$, then the huge values of its estimate are balanced by the large negative values of $\hat{\alpha}_g$.

Since we are not sure when the Markov chain is long enough to obtain reliable results, we tried to calculate estimates from scenario 2 for a longer burn-in period, $B = 15000$ and $B = 30000$. Unfortunately, even this step did not improve the performance of the model. We do not include the table of estimates here, because the results were comparable with the estimates given in the Table 5.3 and Table 5.2.

5.3 Discussion

The results of the simulation study are not too encouraging. It did not perform badly for class 3 and we also got some not entirely nonsensical results for class 2, e.g., it was able to recognize that it is the largest class in the population. Probably, due to the fact that these classes behave in the opposite direction with respect to the longitudinal model and it is easy to distinguish between them. However, class 1 gathers the subjects that do not behave either as class 2 or 3 but are more or less without the effect of a new drug and with a moderate time effect. Then it was not easy for the algorithm to classify subjects correctly into this class and it led to a mixing classes 1 and 2. The estimation process had the biggest problem to classify correctly the subjects to class 1. Moreover, the estimation of the parameters of the Cox model was definitely less precise than for the parameters of the longitudinal model and it required a larger sample size because the estimates in the Cox model were unstable for small samples. However, we cannot be sure that we let Markov chains run long enough because there is no rule for choosing the length of simulated chains. Longer chains could lead to better results.

Furthermore, we have tried only some parameters to be specific to individual classes, perhaps this approach is better for studying populations where, for example, random effects are class-specific. Nevertheless, there was not enough space to consider all the possibilities and evaluate to which specifics of the class this method of estimation is most sensitive.

Next, this simulation study was somewhat limited by the application of JAGS. In several situations, calculations could not be completed due to the JAGS settings. It would be better to develop our own functions (e.g. create a new R package) so that we can calculate everything ourselves because of the drawbacks that JAGS has. However, it was beyond the scope of this work. For future work, we suggest to provide users with an R package, maybe then the application of the suggested methods in the thesis will be more user-friendly, and it would be used more often in data analysis.

Conclusion

In this thesis the joint models for longitudinal and time-to-event data were introduced. We mentioned two types of these models, namely joint latent class models and joint models with shared random effects. The models were shortly presented and due to the fact that there are several issues with these models that can be discussed we focus on one of the problems.

Therefore, the rest of the thesis was focused on the joint latent class models. The main intention of this thesis was to describe a parameter estimation of the model from the Bayesian point of view. This involves a proper definition of the model in the Bayesian framework, then the discussion of suitable choices of prior distributions for all parameters. The crucial role of the Bayesian estimation of the parameters in the model is played by full conditional distributions, which are then used in computational algorithms for parameter estimation. Thus, we derived full conditional distributions for all parameters of interest in the model. For all of the parameters of the longitudinal model we found out that the form of the derived densities was proportional to densities of some standard distributions. The same cannot be said for the derived densities of the parameters related to the Cox model, apart from the derived density of the scale parameter η_g, the densities are not proportional to any form of standard distribution. Last but not least, the full conditional distributions were derived for the parameters of the class probabilities. We mentioned two options of how to define class probabilities. For only one option, where the probability does not depend on any other parameter, the resulting density is proportional to the standard distribution. These results can be used in an algorithm (e.g. the Metropolis-Hasting or the Gibbs algorithm) to compute the estimates of the model parameters.

We did not provide the reader with our own functions for computing the estimates, however we defined the model using `JAGS` software and performed a small simulation study. The results of the study were not optimal, nevertheless there are many other settings that were not examined and the Bayesian methods can work for the different setting much better. We discussed the drawbacks of the usage of `JAGS` and for future work we propose to develop a new functions or `R` package, that would simplify the application of the Bayesian estimation methods for users.

Moreover, there are a large number of problems that can be discussed in more detail for the joint models or specifically for the latent class joint models. For instance, to our best knowledge, there is no literature that focuses on selecting the number of latent classes in the latent class joint model when using the Bayesian approach to compute the estimates. However, covering all of these issues was beyond the scope of this work.

Bibliography

D. Alvares, C. Armero, and A. Forte. What does objective mean in a dirichlet-multinomial process? *International Statistical Review*, 86(1):106–118, 2018.

J. Anděl. *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha, 2007. ISBN 80-7378-001-1.

E.-R. Andrinopoulou, K. Nasserinejad, R. Szczesniak, and D. Rizopoulos. Integrating latent classes in the bayesian shared parameter joint model of longitudinal and survival outcomes. 2018.

S. Brooks, A. Gelman, G. Jones, and X. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

D. R. Cox. Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, 34(2):187–220, 1972.

P. J. Diggle, I. Sousa, and A. G. Chetwynd. Joint modelling of repeated measurements and time-to-event outcomes: The fourth armitage lecture. *Statistics in Medicine*, 27(16):2981–2998, 2008.

Ch. L. Faucett and D. C. Thomas. Simultaneously modeling censored survival data and repeatedly measured covariates: A gibbs sampling approach. *Statistics in Medicine*, 15(15):1663–1685, 1996.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

J. Han, E. H. Slate, and E. A. Peña. Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Statistics in medicine*, 26:5285–5302, December 2007. ISSN 0277-6715. doi: 10.1002/sim.2915.

D. A. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61((2)):383–385, 1974.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

C. Henderson. *Application of Linear Models in Animal Breeding*. University of Guelph, 1984.

L. G. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 2016.

R. I. Jennrich and M. D. Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820, 1986.

N. M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963, 1982.

A. Lawrence Gould, M. E. Boye, M. J. Crowther, J. G. Ibrahim, G. Quartey, S. Micallef, and F. Y. Bois. Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in medicine*, 34(14):2181–2195, 2015.

M. J. Lindstrom and D. M. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):370–384, 1988.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

K. Nasserinejad, J. van Rosmalen, W. de Kort, and E. Lesaffre. Comparison of criteria for choosing the number of classes in bayesian finite mixture models. *PloS one*, 12(1), 2017.

E. Njeru Njagi, D. Rizopoulos, G. Molenberghs, P. Dendale, and K. Willekens. A joint survival-longitudinal modelling approach for the dynamic prediction of rehospitalization in telemonitored chronic heart failure patients. *Statistical Modelling*, 13:179–198, 2013.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

M. Plummer. Jags: A program for analysis of bayesian graphical models using gibbs sampling, 2003.

C. Proust-Lima, H. Amieva, and H. Jacqmin-Gadda. Analysis of multivariate mixed longitudinal data: a flexible latent process approach. *British Journal of Mathematical and Statistical Psychology*, 66(3):470–487, 2013.

C. Proust-Lima, M. Séne, J. Taylor, and H. Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, 23(1):74–90, 2014.

C. Proust-Lima, V. Philipps, and B. Liquet. Estimation of extended mixed models using latent classes and latent processes: the r package lcmm. 2015. doi: 10.18637/jss.v078.i02.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

D Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Boca Raton: CRC Press, 2012.

Ch. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

M. S. Wulfsohn and A.A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339, 1997. ISSN 0006341X, 15410420.

A. Attachments

A.1 Probability distributions

In this part of the Appendix we recapitulate definitions of some of the standard distributions used in the thesis. The main purpose is that in some cases several parametrizations are used and we would like to make it clear which parametrization was used in the text.

Definition. The random variable $\mathbb{X}$ follows a Wishart distribution with a density of the form

$$f(\mathbb{X}) = \Big\{2^{\frac{qd}{2}}\pi^{\frac{q(q-1)}{4}}\prod_{i=1}^{q}\Gamma\Big(\frac{d+1-i}{2}\Big)\Big\}|\mathbb{B}|^{-\frac{d}{2}}|\mathbb{X}|^{\frac{d-q-1}{2}}\exp\{-\frac{1}{2}\mathrm{tr}(\mathbb{B}^{-1}\mathbb{X})\}, \quad \mathbb{X} > 0$$

where $\mathbb{B}$ is a positive definite matrix, $d > q-1$ are degrees of freedom, $\mathbb{X}$ is a positive definite $(d \times d)$-matrix, and we write $\mathbb{X} \sim \mathsf{W}_q(d, \mathbb{B})$.

Definition. The random variable X follows a Weibull distribution with a density of the form

$$f(x) = \eta\nu x^{\nu-1}\exp\{-\eta x^{\nu}\}, \quad x \geq 0$$

where $\nu > 0$ and $\eta > 0$ and we write $X \sim \mathsf{We}(\nu, \eta)$.

Definition. The random variable $\boldsymbol{X}$ follows a Dirichlet distribution with a density of the form

$$f(\boldsymbol{x}) = \frac{1}{B(\boldsymbol{a})}\prod_{i=1}^{K}x_i^{a_i-1}, \quad \text{where } B(\boldsymbol{a}) = \frac{\prod_{i=1}^{K}\Gamma(a_i)}{\Gamma(\sum_{i=1}^{K}a_i)}, \sum_{i=1}^{K}x_i = 1, x_i \in (0,1),$$

where $X = (X_1, \ldots, X_K)^{\mathsf{T}}$, $\boldsymbol{a} = (a_1, \ldots, a_K)^{\mathsf{T}}, a_i > 0$ and we write $\boldsymbol{X} \sim \mathsf{Dir}(\boldsymbol{a})$.

A.2 Matrix Algebra

In the following theorem we put together the properties of the trace of the matrix that are used in the derivation of the full conditional distributions. This knowledge comes from the textbooks of linear algebra.

Theorem 2. *Suppose that $\mathbb{A}_i$ are $(n \times n)$-matrices for $i = 1, \ldots, K$, $\mathbb{B}$ is an $(n \times m)$-matrix, $\mathbb{C}$ is an $(m \times n)$-matrix, and $c \in \mathbb{R}$, then the following holds*

(i) $tr(\sum_{i=1}^{K} c\mathbb{A}_i) = c\sum_{i=1}^{K} tr(\mathbb{A}_i)$,

(ii) $tr(\mathbb{B}\mathbb{C}) = tr(\mathbb{C}\mathbb{D})$.