FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Jan Vávra

Model-based Clustering of Multivariate Longitudinal Data of a Mixed Type

Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D.
Study programme: Probability and statistics, econometrics
and financial mathematics

Prague 2022

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

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 ....

Author's signature

First of all, I would like to thank to my supervisor doc. RNDr. Arnošt Komárek, Ph.D. for his guidance in the world of research, for his expert counsel, for his invaluable advice, for his time and patience. To Vladislav Bína I owe my thanks for introducing me to the EU-SILC database. An integral part of my later research was based on the work of Bettina Grün and Gertraud Malsiner-Walli, who were so kind and invited me to spend several weeks at Vienna University of Economics and Business. The dissemination of our research at several national and international conferences would not have been possible without the financial support of the Czech Science Foundation (GAČR) grant 19-00015S and the Charles University, GA UK project no. 298120.

Last but not the least, countless thanks to my family and friends who have cared for my physical and mental health during my studies, spanning the dark ages of the COVID-19 pandemics, and supported me in every other conceivable way that would be impossible to list without leaving something out.

Title: Model-based Clustering of Multivariate Longitudinal Data of a Mixed Type

Author: Jan Vávra

Department: Department of Probability and Mathematical Statistics

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

Abstract: In many nowadays studies, the data are collected repeatedly on the same units over a certain period of time. Moreover, such longitudinal data are composed of numeric values, count variables, binary indicators, ordered or nominal categories. A few variants of statistical model capable of modelling such often highly correlated data jointly are introduced. On top of that, a methodology of model-based clustering is adapted to such models to discover hidden heterogeneity within the data by dividing units into clusters of specific characteristics. Bayesian approach is taken, generative model is proposed and MCMC methodology is developed for estimation. A simulation study verifying the estimation properties is conducted. The methodology is applied to real datasets such as medical data on patients suffering from primary biliary cholangitis (PBC) or economical dataset consisting of thousands of Czech households followed since 2005 (EU-SILC database).

Keywords: model-based clustering, MCMC, longitudinal data, GLMM, mixed type

Contents

Notation

Probability distribution characteristics:

P	probability measure
$p(\boldsymbol{x}\vert\boldsymbol{y})$	probability density function (pdf) of $\boldsymbol{X}\vert\boldsymbol{Y}$
$\ell(\boldsymbol{x}\vert\boldsymbol{y})$	logarithm of pdf of $\boldsymbol{X}\vert\boldsymbol{Y}$
$\boldsymbol{X}\vert\cdots$	full-conditional distribution of $\boldsymbol{X}$, i.e. distribution of $\boldsymbol{X}$ given all other random elements
$\mathsf{E}\,\boldsymbol{X}$	expected value of a random vector $\boldsymbol{X}$
$\mathsf{var}\,\boldsymbol{X}$	variance matrix of a random vector $\boldsymbol{X}$

Families of probability distributions:

$\mathsf{D}_{\{x\}}$	Dirac degenerate distribution at x
$\mathsf{N}\left(\mu,\,\sigma^2\right)$	normal distribution with mean μ and variance $\sigma^2 > 0$
$\varphi(y;\mu,\sigma^2)$	probability density function of $\mathsf{N}\left(\mu,\,\sigma^2\right)$
$\mathsf{N}_k\left(\boldsymbol{\mu},\,\boldsymbol{\Sigma}\right)$	k-variate normal distribution with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma} \geq 0$
$\varphi(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$	probability density function of $\mathsf{N}_k\left(\boldsymbol{\mu},\,\boldsymbol{\Sigma}\right)$
$\mathsf{TN}\left(\mu,\,\sigma^2,\,a,\,b\right)$	univariate normal distribution with mean μ and variance $\sigma^2 >$ truncated to interval $(a, b) \subset [-\infty, \infty]$
$\mathsf{Unif}\left(a, b\right)$	uniform distribution on an open finite interval (a, b)
$\mathsf{Unif}\left\{a_1, \dots, a_n\right\}$	uniform distribution on a set $\{a_1, \dots, a_n\}$
$\mathsf{Beta}\left(\alpha, \beta\right)$	Beta distribution with parameters $\alpha > 0$ and $\beta > 0$
$\mathsf{Dir}_G\left(\boldsymbol{\alpha}\right)$	G-dimensional Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_G)^\top, \alpha_g > 0$
$\mathsf{Pois}\left(\lambda\right)$	Poisson distribution with parameter $\lambda > 0$
$\mathsf{Bernoulli}\left(p\right)$	Bernoulli trial with probability of success $p \in (0, 1)$
$\mathsf{Bi}\left(n, p\right)$	binomial distribution – n independent $\mathsf{Bernoulli}\left(p\right)$ trials
$\mathsf{Mult}_G\left(n, \boldsymbol{p}\right)$	G-variate multinomial distribution of size n and probabilities $\boldsymbol{p} = (p_1, \dots, p_G)^\top$, $p_1 + \cdots + p_G = 1, p_g > 0$
$\Gamma\left(\alpha, \beta\right)$	Gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$, $\mathsf{E}\,\Gamma\left(\alpha, \beta\right) = \frac{\alpha}{\beta}$
$\mathsf{W}_d\left(\mathbb{S}, \nu\right)$	Wishart distribution of dimension d with scale matrix $\mathbb{S} > 0$ and $\nu > d - 1$ degrees of freedom, $\mathsf{E}\,\mathsf{W}_d\left(\mathbb{S}, \nu\right) = n\mathbb{S}$
$\mathsf{IW}_d\left(\mathbb{S}, \nu\right)$	inverse Wishart distribution of dimension d with scale matrix $\mathbb{S} > 0$ and $\nu > d - 1$ degrees of freedom, $\boldsymbol{\Sigma} \sim \mathsf{IW}_d\left(\mathbb{S}, \nu\right) \Leftrightarrow \boldsymbol{\Sigma}^{-1} \sim \mathsf{W}_d\left(\mathbb{S}, \nu\right)$

Special symbols:

$\mathsf{diag}(a_1, \dots, a_n)$	a diagonal matrix with elements a_i on the diagonal
$\mathsf{diag}(\mathbb{A})$	the diagonal of a matrix $\mathbb{A}$
$\lvert\mathbb{A}\rvert$	determinant of a matrix $\mathbb{A}$
$\mathsf{Tr}(\mathbb{A})$	trace of a matrix $\mathbb{A}$
$\boldsymbol{a}^\top, \mathbb{A}^\top$	transposition of a vector $\boldsymbol{a}$ and of a matrix $\mathbb{A}$
$\mathbb{A}^{-1}$	inverse matrix to a matrix $\mathbb{A}$
$\mathbb{A} > 0$	positive-definite matrix $\mathbb{A}$
$\mathbb{A} \geq 0$	positive semi-definite matrix $\mathbb{A}$
$\mathbf{0}_n$	n-long vector of zeros
$\mathbb{O}_{n \times m}$	$n \times m$ zero matrix
$\mathbf{1}_n$	n-long vector of ones
$\mathbb{I}_n$	unit matrix of order n, $\mathsf{diag}(\mathbf{1}_n)$
$\mathbb{1}_{\mathcal{A}}(a) = \mathbb{1}_{(a \in \mathcal{A})}$	indicator function, 1 if $a \in \mathcal{A}$, 0 otherwise
$\overset{!}{=}$	find a solution of a system of equations
$:=$	define
$+=$	add the following value to the current one
$-=$	subtract the following value from the current one
$*=$	multiply the current value by the following one
$/=$	divide the current value by the following one
$m:n$	$\{m, m+1, \dots, n\}$ in algorithmic notation
$\Gamma(\cdot)$	gamma function
$\mathsf{B}\left(\alpha, \beta\right)$	beta function

List of acronyms

ACF	autocorrelation function
AIC	Akaike information criterion
AGQ	adaptive Gaussian quadrature
BDA	Bayesian data augmentation
BIC	Bayesian information criterion
BUGS	Bayesian inference using Gibbs sampling
DIC	deviance information criterion
ECDF	empirical cumulative distribution function
EM	expectation-maximization (algorithm)
ET	equal-tailed (credible intervals)
EU-SILC	European Union – Statistics on Income and Living Conditions
GLMM	generalized linear mixed-effects model
HPD	highest posterior density (credible intervals)
JAGS	just another Gibbs sampler
LME	linear mixed-effects (model)
MBC	model-based clustering
MCMC	Markov chain Monte Carlo
MLE	maximum likelihood estimator
PBC	primary biliary cholangitis
pdf	probability density function
SLLN	strong law of large numbers
wrt	with respect to

Introduction

Scientific progress in the information era is based on the constant collection of data. As time passes, medical studies, population surveys, etc. become more sophisticated, collect more and more data of diverse nature or even follow the same units over longer periods of time. However, if we are to learn from them in the future, the development of statistical methods for analysing these complex datasets must keep pace.

Within this thesis we will be particularly interested in longitudinal (panel) data gathered repeatedly on the same units over time. Though, the units could be considered independent among themselves, the several observations from the same unit cannot. Moreover, the data come in diverse forms: from categories of binary, ordinal or nominal nature, through count outcomes, to specific numeric values. Plenty univariate models for treating such types of outcomes have already been proposed. Among the most favourite univariate models for longitudinal data are the generalized linear mixed-effects models, which can adapt to any type of outcome (Laird and Ware, 1982; Stiratelli et al., 1984; Jiang, 2007). However, joint modelling of potentially highly correlated outcomes of different type evades broad attention. The general idea for combining random-effects models was provided by Fieuws and Verbeke (2004).

Even though some reasonable statistical model for multivariate longitudinal mixed-type data is found, the reality is often much more complicated. Close explorative analysis may reveal that the reason why many of the studied units do not fit the supposed general trend is that they belong to different subgroups of their own characteristics. Luckily, Banfield and Raftery (1993) already proposed a methodology for capturing the different patterns and giving an order to such heterogeneous data instead of averaging them together. Although they introduced the model-based clustering methodology only for Gaussian mixtures, it was only a matter of time before it would be applied to more complex systems such as the longitudinal data (Molenberghs and Verbeke, 2005). Nevertheless, there is less improvement in the clustering of several longitudinal outcomes jointly that would also provide a ready to use software implementation. Though, several options are available, they do not fulfil all our demands. Proust-Lima et al. (2017) model only outcomes of the same type. Grün and Leisch (2008) cluster mixed-type data under the independence of the outcomes, which is far from the real world experience. The solution to the outlined problem that is the closest to our ideas and imagination is the work of Komárek and Komárková (2013) who jointly model numeric, binary and count outcome.

Nevertheless, our ambitions for the model and its capabilities are even higher. We would like to cover any combination of several possibly highly correlated numeric, count, binary, ordinal and general categorical outcomes in one single model. Moreover, we would like to give future analysts a freedom in specification of the form of heterogeneity within the data depending on their expectations. During our research we even decided to address several issues such as the choice of the number of underlying groups or missing outcome values. In four years of our research we developed such a model and provided our own implementation in the free statistical software R (R Core Team, 2022).

This thesis mainly consists of two publications in impacted journals:

- J. Vávra and A. Komárek. Classification based on multivariate mixed type longitudinal data: With an application to the EU-SILC database. *Advances in Data Analysis and Classification*, 2022. doi: https://doi.org/10.1007/s11634-022-00504-8.

  This paper presents an initial statistical model for clustering multivariate longitudinal data of mixed type. Binary and ordinal outcomes are modelled by thresholded latent numeric outcomes. Multivariate linear mixed-effects model is supposed for all numeric outcomes (both observed and latent). Model-based clustering approach is applied to allow for different evolution patterns to capture hidden heterogeneity within the data.

- J. Vávra, A. Komárek, B. Grün, and G. Malsiner-Walli. Clusterwise multivariate regression of mixed-type panel data. *Submitted, available as preprint.* doi: https://doi.org/10.21203/rs.3.rs-1882841/v1.

  The main objective of this follow-up paper was to allow for more outcome types. Therefore, the threshold concept model was replaced by generalized linear mixed-effects models where a large number of combinations of distributional families and link functions are potentially available. Logistic regression, ordinal logit regression and multinomial regression were used for binary, ordinal and general categorical outcomes, respectively. The implementation also allowed to overcome the issue of missing outcome values. In addition, the problem of a priori unknown number of mixture components was solved by the sparse finite mixtures – a methodology originally developed by our co-authors Bettina Grün and Gertraud Malsiner-Walli.

Progress in our research, including real data analyses, have been presented at several international conferences. Some contributions have even been included within the conference proceedings:

- J. Vávra and A. Komárek. Identification of Temporal Patterns in Income and Living Conditions of Czech Households: Clustering Based on Mixed Type Panel Data from the EU-SILC Database. *38th International Conference on Mathematical Methods in Economics, Conference Proceedings*, 612–617, 2020.

- J. Vávra and A. Komárek. Clustering Based on Multivariate Mixed Type Longitudinal Data with an application to the EU-SILC database. *Proceedings of the 22nd European Young Statisticians Meeting*, 148–152, 2021.

- J. Vávra. GLMM Based Segmentation of Czech Households Using the EU-SILC Database. *Proceedings of the 39th International Conference on Mathematical Methods in Economics*, 505–510, 2021.

- J. Vávra and A. Komárek. GLMM Based Clustering of Multivariate Mixed Type Longitudinal Data. *Proceedings of the 36th International Workshop on Statistical Modelling*, 337–342, 2022.

This thesis compiles the listed publications into one clear document of unified notation. Many aspects of the two suggested modelling approaches are analogous, hence, mentioned only once. On the other hand, the major differences are highlighted and compared in different (sub)sections.

The thesis opens with the introduction to the notation used for longitudinal data of a mixed type. Then, real datasets are introduced: well-known PBC dataset which continuously serves as an illustrative example for our modelling techniques and the EU-SILC database. Chapter 2 introduces different approaches for modelling longitudinal data and explains how the highly correlated outcomes are modelled jointly by one multivariate model. Once the statistical models are fully established, we introduce the model-based clustering framework for dividing units into several homogenous groups and show how it is applied to both suggested statistical models in Chapter 3. Estimation and inference is performed in Bayesian fashion, hence, we dedicate Chapter 4 to establish the prior distributions for model parameters and have a quick overview of the posterior distribution. MCMC methods (Gibbs sampling and Metropolis proposals) explore this very complex posterior distribution. The design of the used samplers is theoretically justified for the two models separately in Chapters 5 and 6. These chapters are closed by the most important results of the performed simulation study. Chapter 7 contains details about the implementation including several frequently used algorithms. The thesis concludes with the detailed analysis of the longitudinal data on the living conditions of Czech households from EU-SILC database in Chapter 8.

1. Longitudinal mixed-type data

The *longitudinal* data or often called *panel* data arise when a given set of outcomes is observed regularly on a studied unit. For example, a patient is obliged to regularly visit a doctor for a unified medical examination (a series of blood tests, etc.), a household each year fills a questionnaire of a unified form. *Functional* data satisfy such a condition too, however, under much higher frequency, which makes the data seem as several independent smooth curves. Here, only several observations per observed study unit will be assumed, however, the sample size of the independent units is assumed to be large.

Let us declare the notation precisely. In total, R different outcomes denoted by $Y^r, r = 1, \dots, R$ will be observed. The dataset consists of n independently behaving units. A unit $i, i = 1, \dots, n$, is observed at n_i time points $t_{i,1}, \dots, t_{i,n_i}$ (each unit is allowed to be observed for a different number of times $n_i \in \mathbb{N}$). The following notation will be kept throughout the thesis:

- $Y^r_{i,j}$ – a single observation of an outcome r by a unit i at a time point j,
- $\boldsymbol{Y}^r_i = \{Y^r_{i,j}, j = 1, \dots, n_i\}$ – a vector of all observations of outcome r by a unit i,
- $\boldsymbol{Y}^r = \{Y^r_{i,j}, i = 1, \dots, n,\ j = 1, \dots, n_i\}$ – all observations of outcome r,
- $\mathbb{Y}_i = \{\boldsymbol{Y}^r_i, r = 1, \dots, R\}$ – a collection of all observations of all outcomes by a unit i,
- $\mathbb{Y} = \{\mathbb{Y}_i, i = 1, \dots, n\}$ – a collection of all observed data,

where $i = 1, \dots, n,\ j = 1, \dots, n_i,\ r = 1, \dots, R$. Alongside the outcomes of interest, many other additional variables are recorded and may be used as explanatory variables, especially, the time points $t_{i,j}$ are considered to be a typical covariate. Analogously as before, we denote by

- $\mathcal{C}_{i,j} = \{t_{i,j}, \dots\}$ – the covariate values of a unit i at a time point j,
- $\mathcal{C}_i = \{\mathcal{C}_{i,j}, j = 1, \dots, n_i\}$ – a collection of the covariate values of a unit i from all the time points,
- $\mathcal{C} = \{\mathcal{C}_i, i = 1, \dots, n\}$ – a collection of all the covariate values,

where $i = 1, \dots, n,\ j = 1, \dots, n_i$. Hence, the couple $\{\mathbb{Y}, \mathcal{C}\}$ represents the whole dataset at our disposal that consist of n independent blocks $\{\mathbb{Y}_i, \mathcal{C}_i\}$.

The letters for indices $i = 1, \dots, n,\ j = 1, \dots, n_i$ and $r = 1, \dots, R$ will be reserved for this meaning throughout the whole thesis unless stated otherwise. Hence, we shall simplify the expressions by this convention. Nevertheless, in case the range of possible values has to be restricted, the restriction will be specified. We will often have to restrict the index r based on the type of the corresponding outcome.

1.1 Types of outcomes considered

We assume R outcomes of interest in total, let us denote by $\mathcal{R}$ the index set for these outcomes, i.e. $\mathcal{R} = \{1, \dots, R\}$. Commonly, all R outcomes are of the same type and a multivariate version of statistical model is developed. However, here we focus on *mixed-type* data, that is, when there are groups of outcomes of different natures. Here we list the outcome types and the corresponding notation.

First, a typical and most informative outcome is the *numeric* outcome, distribution of which could be considered continuous, e.g. concentration of a substance of interest in a blood sample or the overall income of the household. The indices corresponding to *numeric* outcomes will be denoted by $\mathcal{R}^{\mathsf{Num}} \subset \mathcal{R}$.

Slightly less informative are *count* outcomes expressing total number of specific instances, e.g. number of events observed in a certain period of time or the platelet count within the blood sample. In some circumstances, there could be an argument for considering such an outcome rather as *numeric*, however, we still distinguish this specific type of outcome because it may demand a special type of model such as Poisson regression. Hence, the indices corresponding to *count* outcomes will be denoted by $\mathcal{R}^{\mathsf{Poi}} \subset \mathcal{R}$.

Next, we have to deal with categorical outcomes, where the recorded numbers correspond to a certain category. The most elementary case is the *binary* type of outcome, where only two levels are distinguished. Usually, level 1 denotes a success in some criterion (e.g. presence of a given medical symptom or affordability of a certain luxury within the household), while level 0 stands for the opposite. The indices corresponding to *binary* outcomes will be denoted by $\mathcal{R}^{\mathsf{Bin}} \subset \mathcal{R}$.

In case of $K^r > 2$ levels, it is beneficial to distinguish an *ordinal* outcome from the *nominal* (*general categorical*) one. The ordered levels $0 < 1 < \dots < K^r - 1$ may represent a scale from the most negative to the most positive outcome, e.g. self-evaluation of given criterion on a scale from 1 to 10. It may even be a result of categorization of a *numeric* outcome. The indices corresponding to *ordinal* outcomes will be denoted by $\mathcal{R}^{\mathsf{Ord}} \subset \mathcal{R}$.

General categorical outcomes of $K^r > 2$ levels without any evident ordering cannot utilize the ordinality for a simpler model and, hence, have to be separated into an individual group. The indices corresponding to *general categorical* outcomes will be denoted by $\mathcal{R}^{\mathsf{Cat}} \subset \mathcal{R}$.

Prior the analysis, the analyst decides, into which one of the sets $\mathcal{R}^{\mathsf{Num}}$, $\mathcal{R}^{\mathsf{Poi}}$, $\mathcal{R}^{\mathsf{Bin}}$, $\mathcal{R}^{\mathsf{Ord}}$, $\mathcal{R}^{\mathsf{Cat}}$ each index $r \in \mathcal{R}$ belongs. Which means that these sets are disjoint and $\mathcal{R}^{\mathsf{Num}} \cup \mathcal{R}^{\mathsf{Poi}} \cup \mathcal{R}^{\mathsf{Bin}} \cup \mathcal{R}^{\mathsf{Ord}} \cup \mathcal{R}^{\mathsf{Cat}} = \mathcal{R}$. Many of the later notations will depend on the type of the considered outcome. In such a case, we abbreviate the types by letters $\{\mathsf{N}, \mathsf{P}, \mathsf{B}, \mathsf{O}, \mathsf{C}\}$ and declare a function declaring a type of an outcome $r \in \mathcal{R}$:

$$\mathsf{t}(r) := \mathsf{type}(r) = \begin{cases} \mathsf{N}, & \text{if } r \in \mathcal{R}^{\mathsf{Num}}, \\ \mathsf{P}, & \text{if } r \in \mathcal{R}^{\mathsf{Poi}}, \\ \mathsf{B}, & \text{if } r \in \mathcal{R}^{\mathsf{Bin}}, \\ \mathsf{O}, & \text{if } r \in \mathcal{R}^{\mathsf{Ord}}, \\ \mathsf{C}, & \text{if } r \in \mathcal{R}^{\mathsf{Cat}}. \end{cases}$$

If each of the types appears at most once, we will specify the type in the indexation by the letters directly, e.g. Y_{ij}^{N} instead of $Y_{ij}^{r}, r \in \mathcal{R}^{\mathsf{Num}}$.

1.2 PBC dataset

This dataset is named after *primary biliary cirrhosis* (PBC) an autoimmune disease slowly leading to liver decompensation. In 2014, it was renamed to *primary biliary cholangitis* (still fitting the abbreviation PBC) to differentiate this autoimmune disease from the cirrhosis which is only a feature of advanced stadium.

This dataset was gathered at the Mayo Clinic between 1974 and 1984 to monitor patients suffering from this disease. Moreover, a case-control study was performed to evaluate the effect of the drug D-penicillamine. Out of 424 eligible patients, only 312 cases consented to and participated in the randomized trial. For these randomized patients a large variety of biomedical markers (outcomes) has been recorded, while the remaining 112 have undergone only basic tests. After the initial medical examination, regular visits were scheduled at 6 months, 1 year, and annually thereafter. In case of worsening the medical conditions, extra visits were undertaken, however, not all tests were performed, which leaves some blank spaces within the dataset.

1.2.1 Goals and restriction to `PBC910`

Researchers (Dickson et al., 1989; Therneau and Grambsch, 2000) aimed to estimate the effect of biomedical markers and other patients characteristics on survival, for which the well known Cox model was used.

Our goals will be slightly different. We want to model directly the chosen and highly related markers (of diverse nature) and discover several different trends, some of which may even indicate worsening of the medical condition of a patient. The discovered groups of similar characteristics could be considered as a prognosis groups. Based on the available data of a newly observed patient, we would classify this patient into one of these prognosis groups and then treated him accordingly.

Such an approach was already used by Komárek and Komárková (2013), where they limited the analysis to the dataset `PBC910` available in R package `mixAK` (Komárek and Komárková, 2014). `PBC910` consists of only $n = 260$ patients still alive (without any liver transplantation) after 910 days (2.5 years). Only the data observed within this 2.5 year-long period were used to build a statistical model capable of classifying any other patients with measurements from this initial period. The vast majority (178) of patients have $n_i = 4$ visits recorded within this period. However, there are also patients included where only a single visit is available.

We will follow the steps of Komárek and Komárková (2013) and perform a similar analysis under an extended statistical model. Although, we shall start with the original `pbcseq` dataset (library `survival`) since it offers wider variety of interesting outcomes than `PBC910`, the notation `PBC910` will be kept throughout the thesis.

1.2.2 Outcomes of interest and data summary

Laboratory examinations provide several numeric outcomes regarding concentrations of certain substances in a blood sample, e.g. albumin, alkaline phosphotase, cholesterol, etc. However, the primary numeric outcome for our analysis will be *serum bilirubin* [mg/dl] (`bili`) proven to be associated with survival by a Cox model. To normalize the values we use logarithmic transformation.

The variable *platelet count* (`platelet`) declaring the number of platelets per ml^3/1000 in a span of several tens to hundreds could be considered both count and numeric outcome with a slightly skewed distribution.

There are also a few binary indicators: *presence of hepatomegaly or enlarged liver* (`hepato`), *presence of ascites* or *presence of blood vessel malformations in the skin*, but only the first one will be used throughout the thesis. The `stage` variable is a perfect ordinal variable, however, requires a biopsy to determine the histologic stage of the disease which may not be available for newly observed patient. Hence, we use the other available ordinal outcome `edema` declaring the seriousness of edema by 0 (no edema), 0.5 (untreated or successfully treated edema) or 1 (edema despite diuretic therapy).

Figure 1.1 displays longitudinal profiles of $n = 260$ patients in first 910 days who are still alive at the end of this period. The dark blue patient (`id` $= 220$) has increasing trend in bilirubin concentration as well as decreasing platelet count,

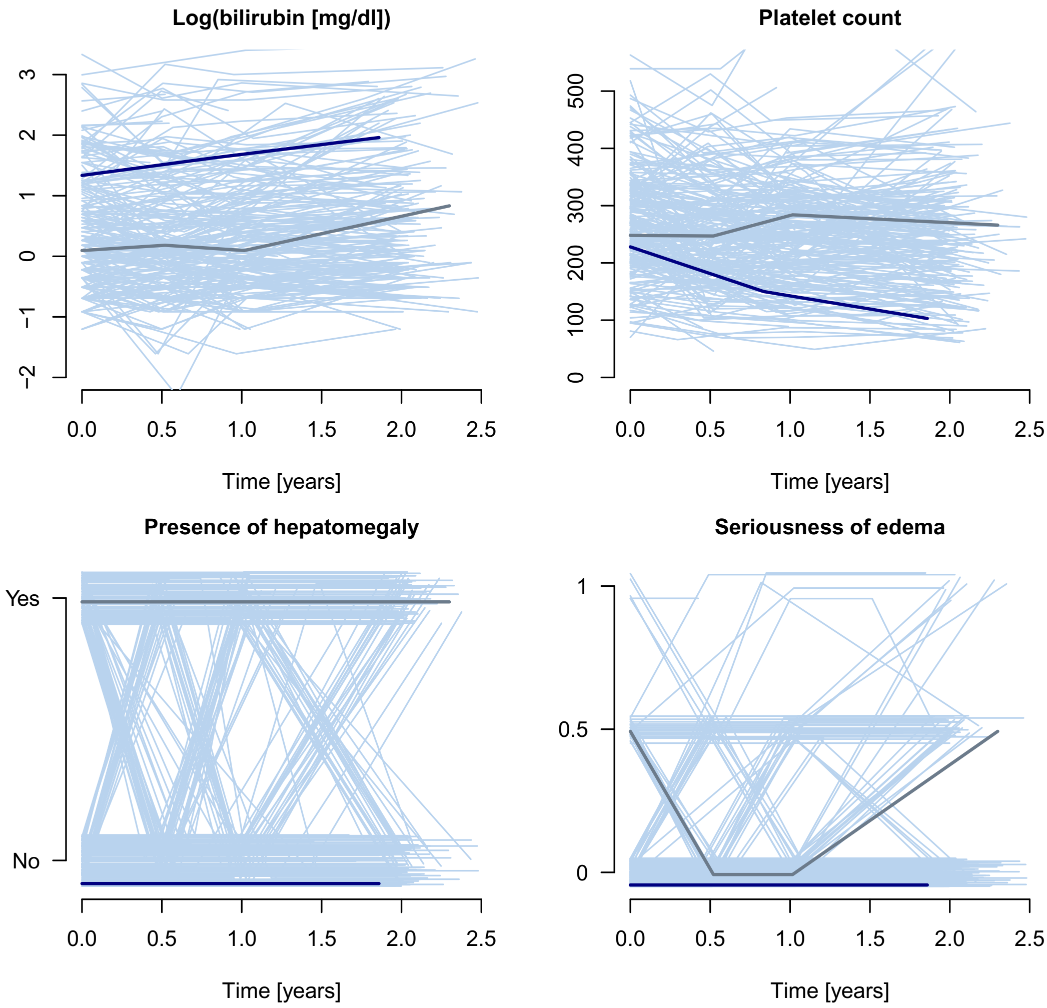


Figure 1.1: `PBC910` longitudinal dataset, two patients highlighted. Outcomes of interest: serum bilirubin on log-scale (numeric), platelet count (count), presence of hepatomegaly (binary) and seriousness of edema (ordinal).

which may be a sign of worsening medical condition compared to the grey patient (`id` = 263). The slight random shifts of the profiles of categorical outcomes reveal not only almost perfect timing of the visits (with respect to the study design), but also the sparsity of severe edema cases (23 out of 918 observations). Presence of hepatomegaly is almost evenly distributed (45.39%).

Apart from evolution in time, we should also consider the effect of age (at the entry to the study) and sex. In Figure 1.2, we plot the averaged values for each patient against the age while distinguishing the gender. We immediately notice the low frequency of males (27 out of 260) within the dataset as well as hints of the trend with increasing age possibly different between the sexes. We should also consider the possible effect of randomization to placebo (125 out of 260) and D-penicillamine drug users (135 out of 260).

Finally, we have to address the relationships among the outcomes themselves. Raw t-tests reveal significant differences in *serum bilirubin* and *platelet count* with respect to different levels of binary and ordinal outcomes. The negative value of Pearson's correlation coefficient (−0.11) between the two numeric outcomes confirms a subtle association between the two outcomes.

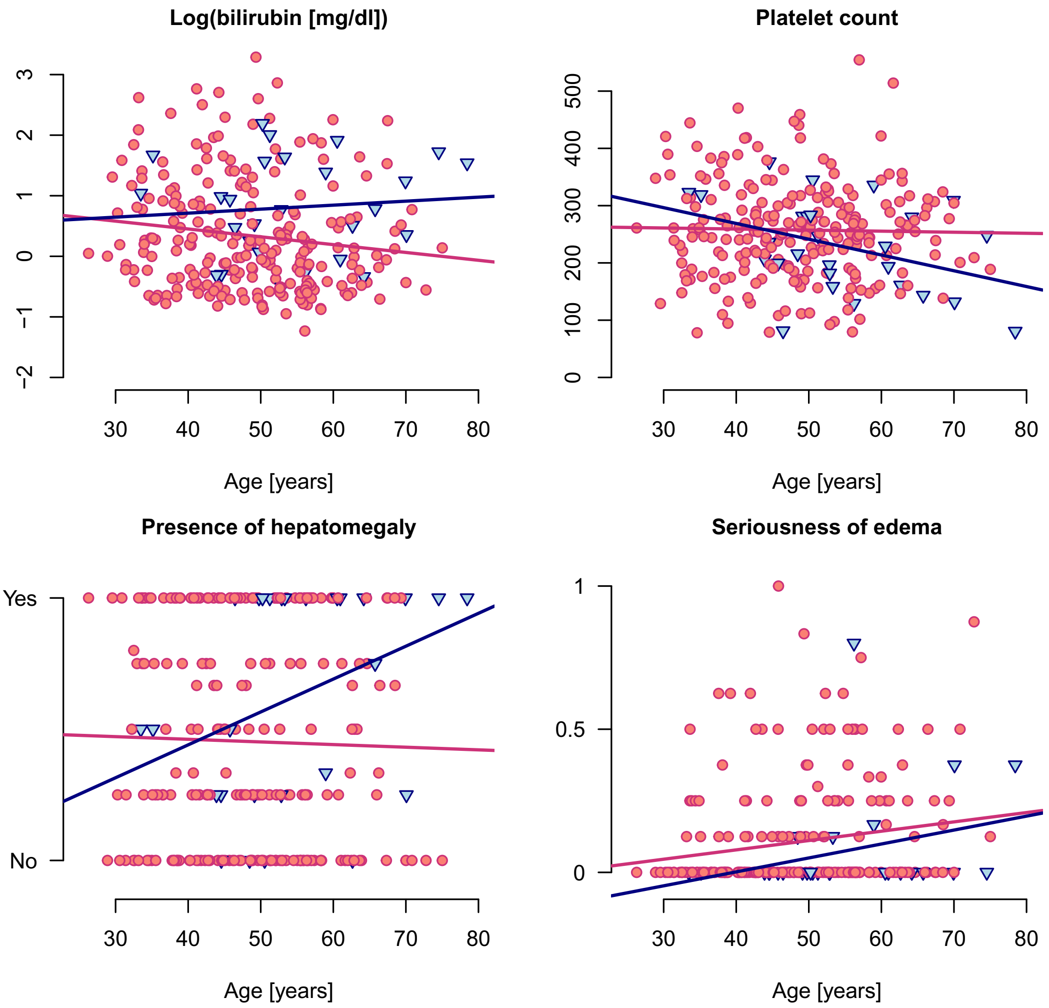


Figure 1.2: `PBC910` longitudinal dataset. Outcomes of interest averaged for each patient; displayed with respect to age and sex (blue for males, red for females). Bold lines are simple linear regression estimates.

1.3 EU-SILC database

The European Union Statistics on Income and Living Conditions database (EU-SILC) is yet another example of longitudinal dataset with outcomes of diverse nature. This still ongoing instrument was launched in 2003 by core members of the European Union with the goal to collect timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. Soon all members of the EU and other European countries including Iceland, Norway and Switzerland agreed to participate in this project. The reference population includes all private households of the respective countries and outcomes which are collected annually via questionnaires.

Though, a comparison of living conditions across all European states is one of the main goals, it is rather ambitious in the scope of our research. Here we will limit ourselves to the subset of Czech households observed from 2005 to 2020.

1.3.1 Design of the study

In general, there are two types of data:

- *cross-sectional data* pertaining to a given time or a certain time period with variables on income, poverty, social exclusion and other living conditions;
- *longitudinal data* pertaining to individual-level changes over time, observed periodically over a four-year period.

Though, they seem to be different, the cross-sectional dataset covers the households within the longitudinal dataset. Hence, if some outcome is not covered by the longitudinal data, one can find it by pairing a household with the corresponding one within the cross-sectional dataset which is in general larger. Moreover, each year an independent module focusing on different aspects (material deprivation, health, . . .) is added. By the nature of our research, our interest lies primarily in the longitudinal dataset covering the same primary outcomes throughout the whole time span.

Each year, the responsible authority (Czech Statistical Office) is obliged to update the set of interviewed households, which is induced by so called *rotational* design illustrated in Figure 1.3 when the system is fully established. As is sketched

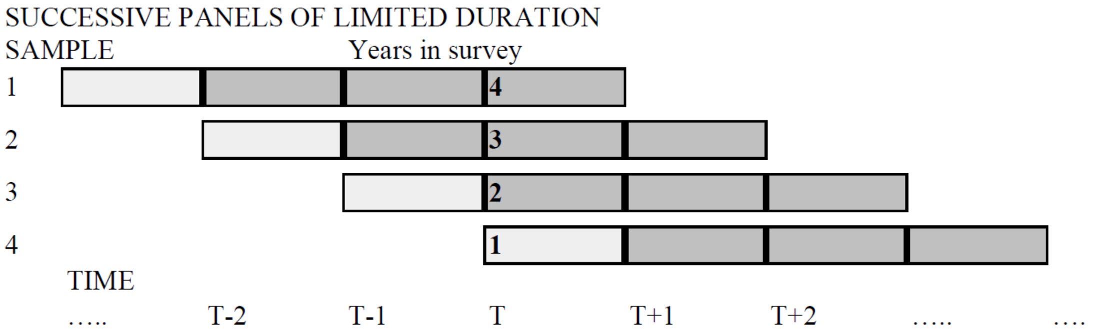


Figure 1.3: Illustration of a simple rotational design once fully established; taken from the official EU-SILC Methodological Guidelines available at (EUS).

in Figure 1.4, the study had to be started by 4 sub-panels: first only for cross-sectional purposes, the second sub-panel is requested to participate only for two years, the third for three years and finally the fourth one for the planned duration of $n_i = 4$ years. In the Czech Republic, more than 7 000 households participated in the first year of the study. Then, each year a quarter of households (the oldest panel) is dropped to be replaced by a set of completely different households of comparable size. By this annual process, it is guaranteed that each household is observed for exactly $n_i = 4$ consecutive years with exception of negligible percentage of households lost during the follow-up for diverse reasons. In total, $n = 27\,386$ Czech households were observed for exactly $n_i = 4$ years between 2005 and 2020.

The data are collected via questionnaires filled during an interview with an adult respondent representing the household. Many variables are measured at a household level which will also be our reference sample unit in our statistical model later. However, wide variety of variables regarding income, education, basic labour information are measured at a personal level. For simplicity of our model, we avoid nested sample units by aggregation to household-level variables.

1.3.2 Outcomes of interest

The list below contains outcomes of interest grouped by the corresponding type:

- Numeric outcomes
  - `HX090` – *Equivalised total disposable income* [EUR/year]

    The sum of gross personal income components (cash, benefits, allowances, rental income, interests, ...) of all household members minus regular taxes (on wealth, income), inter-household cash transfers and social insurance; all divided by the *Equivalised household size*.

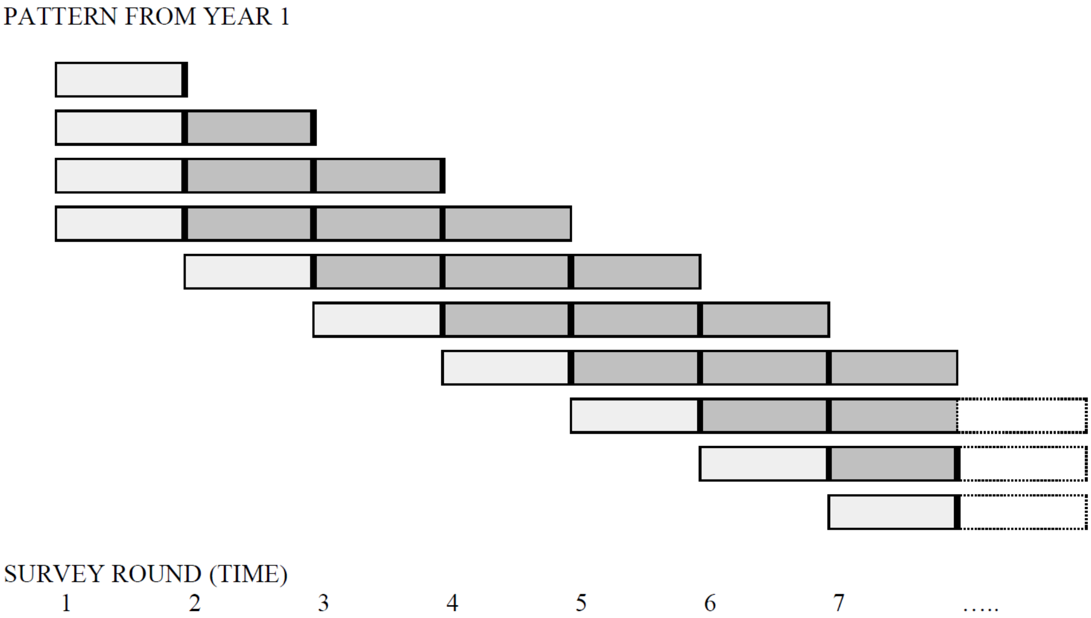


Figure 1.4: Illustration of a simple rotational design in the first years of the study; taken from the official EU-SILC Methodological Guidelines available at (EUS).

  - HS130 – *Lowest income to make ends meet* [EUR/month]

    Respondent's self-assessed indication of the very lowest net monthly income that the household would have to have in order to *make ends meet*, that is to pay its usual necessary expenses (properly defined when asked).

- Binary outcomes (Yes / No)

  - HS040 – *Affordability of a one week holiday*

    Capacity to afford paying for a one week (7 days) annual holiday away from home (yes, if the whole household can afford it regardless of whether the household wants it).

  - HS060 – *Afford to pay for unexpected expenses*

    Capacity to face unexpected financial expenses (surgery, a funeral, major repairs in the house, etc. totalling the equivalent of 1/12th of the national at-risk-of-poverty threshold), that is to pay the expenses through its own resources without taking any loan.

- Ordinal outcomes

  - HS120 – *Ability to make ends meet*

    Thinking of the household's total income, respondent feels to be able to make ends meet (to pay for its usual necessary expenses):

    1. with great difficulty,
    2. with difficulty,
    3. with some difficulty,
    4. fairly easily,
    5. easily,
    6. very easily.

  - HS140 – *Financial burden of the total housing cost*

    The respondent finds the total housing costs including mortgage repayment (instalment and interest) or rent, insurance and service charges (sewage removal, refuse removal, regular maintenance, repairs and other charges):

    1. a heavy financial burden,
    2. a slight financial burden,
    3. not a financial burden at all.

- Categorical outcomes (Yes / No – cannot afford / No – other reason)

  - HS090 – *Do you have a computer?*
  - HS110 – *Do you have a car?*

More details about the definition of these variables is provided by the official EU-SILC Methodological Guidelines available on-line at (EUS).

In applications, we rather work with log-transformed income variables, where a few of negative disposable income values are replaced by 0. The answer to binary outcomes should be *Yes*, if the household can afford a one week holiday or to pay for unexpected expenses despite not doing so in the last year. Ordinal outcomes aim to capture the respondent's feeling about the financial capacities of the household after they have been numerically evaluated. The possession indicators (categorical outcomes) were yes/no questions with the option to specify the reason for not owning the item of interest.

In Figure 1.5 depicting the evolution in time, one can notice a short plateau in the evolution of the outcomes, especially, the *Equivalised total disposable income*. It suggests that many households were impacted by the economical crisis (European sovereign debt crisis) in 2010. Nevertheless, some households may remain untouched by the crisis, which suggests possible heterogeneity within the data. Certainly, we can expect close relations between the outcomes, for example, the higher the disposable income, the higher the chance to afford a week holiday away from home is expected. Close pairwise exploratory analysis confirmed high correlations among the outcomes, which should not be neglected during the statistical modelling.

1.3.3 Covariates

The dataset also offers plenty of covariates that may be related to the outcomes of interest.

- *Time* will be considered as the most important since heterogeneity with respect to evolution in time is expected. We define the time covariate as the number of years past the beginning of 2005, which limits the time into the interval $[0, 16)$. Note that the interviews in the Czech Republic were held in either Q1 or Q2.

- *Equivalised household size* (`HX050`) expresses how large the household is while taking the age of its members into consideration. The head of the household (respondent) has a unit weight, while other members have either 0.5 (older than 14) or 0.3 (younger than 14).

- *Level of urbanisation* was divided by the population density and minimum population into the following categories:

  1. rural – thinly-populated area (non-urban),
  2. town – intermediate area (at least 300 inhabitants per km^2, minimal population of 5 000),
  3. city – densely populated area (at least 1500 inhabitants per km^2, minimal population of 50 000),
  4. Prague – the highly-populated capital city.

  The fourth category was additionally created knowing that the capital city is in many aspects very distinct from the rest of the republic, see for example Figure 1.6. Any other regional (`DB040` – NUTS 2 statistical regions) effects are neglected.

- *The highest ISCED (education) level achieved* within the whole household rarely attains the lowest possible option of primary education. Hence, we merge it with lower-secondary education (label ”Lower”). Then, follows the most common upper-secondary education (label ”Secondary”). Finally, the third category contains both post-secondary and the tertiary education level with a university degree (label ”Higher”).

- *Presence of student* or *baby* indicate whether some household member currently attends any educational institution or is younger than 3 years, respectively.

- *Dwelling type* (`HH010`) classifies households based on the building it lives in. There are distinguished the following types:

    1. detached house,
    2. semi-detached or terraced house,
    3. apartment or flat in a building with less than 10 dwellings,
    4. apartment or flat in a building with 10 or more dwellings,
    5. some other kind of accommodation.

- *Household type* (`HX060`) classifies according to the age composition and role each member has:

    5 - one person household,
    6 - 2 adults, no dependent children, both adults under 65 years,
    7 - 2 adults, no dependent children, at least one adult 65 years or more,
    8 - other households without dependent children,
    9 - single parent household, one or more dependent children,
    10 - 2 adults, one dependent child,
    11 - 2 adults, two dependent children,
    12 - 2 adults, three or more dependent children,
    13 - other households with dependent children,

    where the term *dependent children* is defined as:

    - household members aged 17 or less,
    - economically inactive household members aged between 18 and 24 living with at least one parent.

These covariates will be used to form the predictor of our regression models since they have potentially very important effect on the observed outcomes as Figure 1.7 suggests.

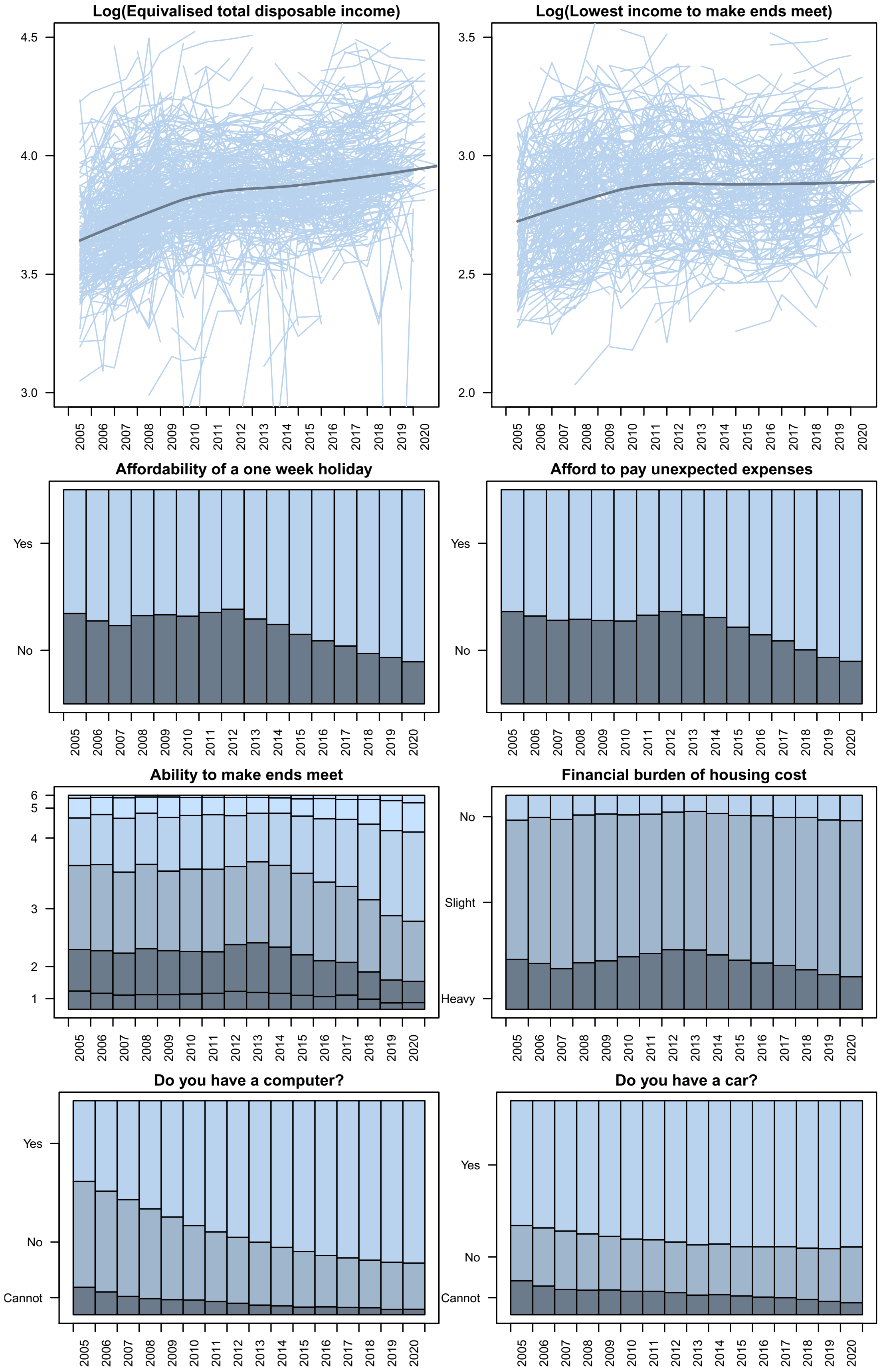
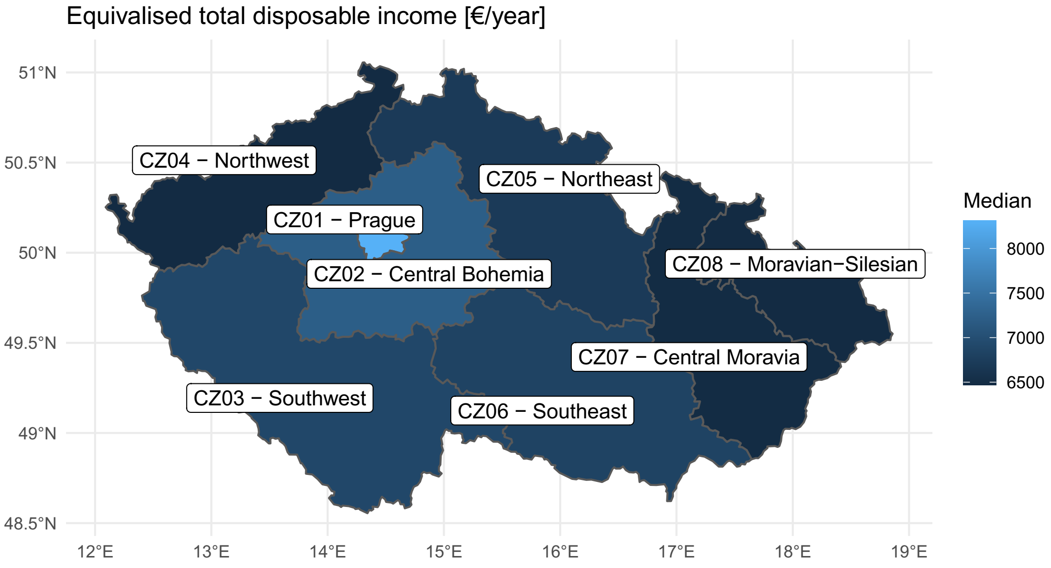
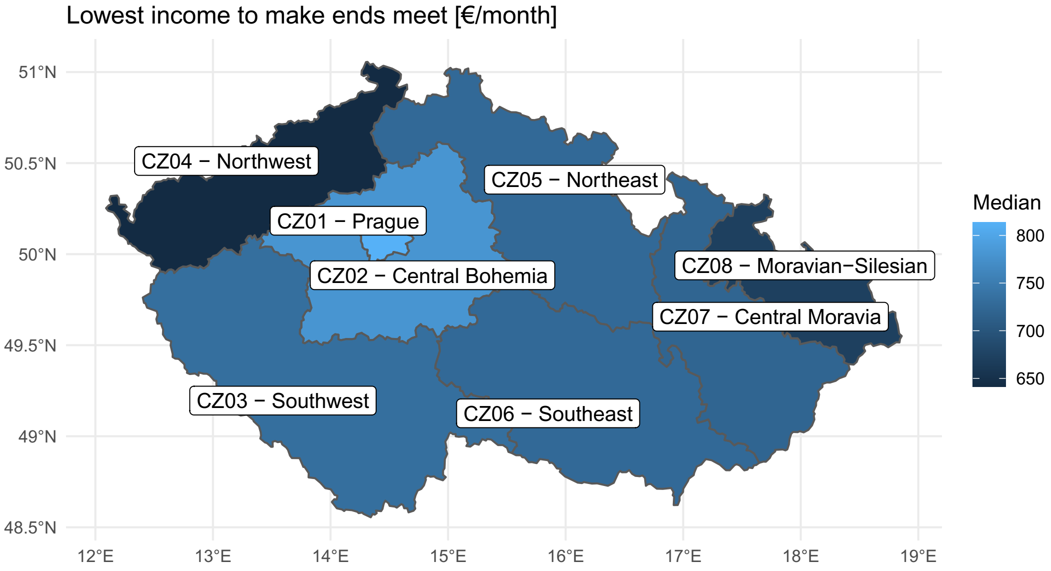


Figure 1.5: EU-SILC dataset. Evolution of outcomes of interest in time. Numeric outcomes are accompanied by a lowess smoothed curve. Level proportions in each year are depicted for categorical outcomes.



(a) *Equivalised total disposable income.*



(b) *Lowest income to make ends meet.*

Figure 1.6: EU-SILC dataset. Median characteristics by NUTS 2 statistical regions of the Czech Republic.

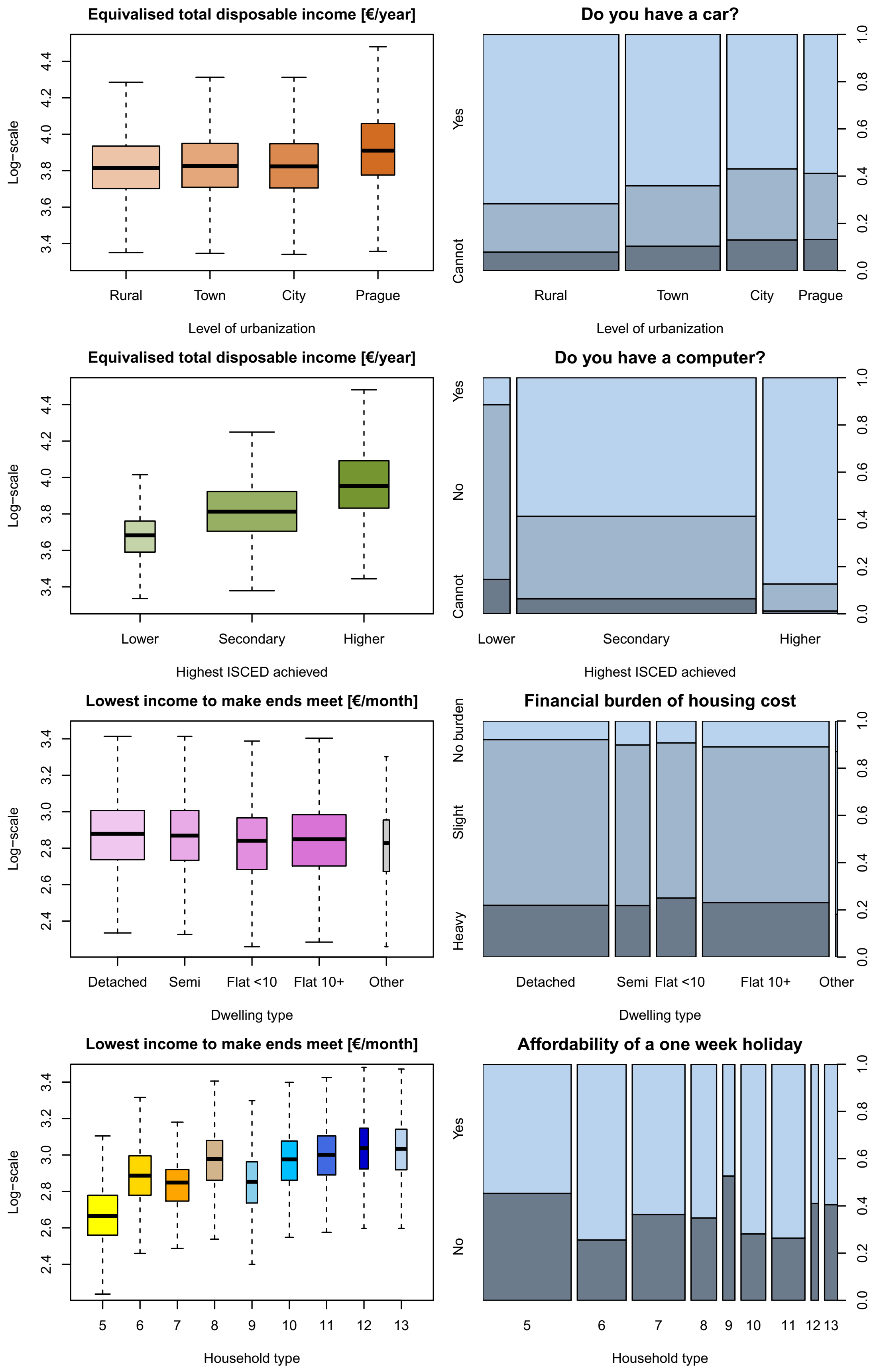


Figure 1.7: EU-SILC dataset. Distribution of chosen outcomes (numeric – boxplots without outliers, categorical – proportions) with respect to different covariates (*Level of urbanization*, *Highest educational level achieved*, *Dwelling type*, *Household type*).

2. Mixed-effects models

When working with longitudinal data introduced in the previous chapter, one enjoys the independence among the units but has to acknowledge the correlation among observations coming from the same unit. Given unit $i = 1, \ldots, n$, the outcome values $Y_{i,j}^r$, $j = 1, \ldots, n_i$ cannot be considered completely independent for any outcome $r = 1, \ldots, R$. Hence, a model for the whole $\boldsymbol{Y}_i^r$ has to be supposed. Moreover, when outcomes cannot be considered independent either, the outcomes $\mathbb{Y}_i$ have to be modelled all together as a block. Nevertheless, the concept of *random effects* allows us to work under independence and still benefit from it.

Laird and Ware (1982) introduced the so called *classical normal linear mixed-effects model* (LME) for a numeric outcome $r \in \mathcal{R}^{\mathsf{Num}}$. Each of the independent units i has its own set of d_r^{R}-dimensional random effects $\boldsymbol{b}_i^r$ which are assumed to be centred and normally distributed independently of each other, i.e. $\boldsymbol{b}_i^r \overset{\mathsf{iid}}{\sim} \mathsf{N}_{d_r^{\mathsf{R}}}(\mathbf{0}, \boldsymbol{\Sigma}_r)$. These latent random effects represent the unit-specific propensity for outcome r with respect to covariates $\boldsymbol{z}_{i,j}^r$. Nevertheless, they remain unobserved and serve only as a tool to establish the independence structure given random effects:

$$Y_{i,j}^r \,\Big|\, \boldsymbol{b}_i^r; \mathcal{C}_{i,j} \sim \mathsf{N}\left(\eta_{i,j}^r, \sigma_r^2\right), \quad \text{where } \eta_{i,j}^r = \left(\boldsymbol{x}_{i,j}^r\right)^\top \boldsymbol{\beta}_r + \left(\boldsymbol{z}_{i,j}^r\right)^\top \boldsymbol{b}_i^r \tag{2.1}$$

is the linear *predictor* formed out of covariates $\mathcal{C}_{i,j}$ and their *fixed* effects $\boldsymbol{\beta}_r$ and the random effects $\boldsymbol{b}_i$. Later, we will rather work with the precision parameter $0 < \tau_r = \left(\sigma_r^2\right)^{-1}$ than the model error variance $\sigma_r^2 > 0$. To introduce notation used for probability density functions and log-likelihood, we remind the reader that single observation $Y_{i,j}^r$ has pdf

$$p_{\mathsf{N}}\left(Y_{i,j}^r = y \,\Big|\, \boldsymbol{b}_i, \tau_r, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) = \varphi\left(y; \eta_{i,j}^r, \tau_r^{-1}\right) \propto \tau_r^{\frac{1}{2}} \exp\left\{\frac{\tau_r}{2}\left(y - \eta_{i,j}^r\right)^2\right\} \tag{2.2}$$

and contributes to the log-likelihood by

$$\ell_{\mathsf{N}}\left(Y_{i,j}^r = y \,\Big|\, \boldsymbol{b}_i, \tau_r, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) = \text{const.} + \frac{1}{2}\log\left(\tau_r\right) - \frac{1}{2}\tau_r\left(y - \eta_{i,j}^r\right)^2. \tag{2.3}$$

The fixed part of the predictor $\eta_{i,j}^{\mathsf{F},r} = \left(\boldsymbol{x}_{i,j}^r\right)^\top \boldsymbol{\beta}_r$ which is a linear combination of regressors $\boldsymbol{x}_{i,j}^r$ derived from the full covariate information $\mathcal{C}_{i,j}$ with the unknown vector of coefficients $\boldsymbol{\beta}_r$ of dimension d_r^{F}, captures the overall trend. On the other hand, the random part $\eta_{i,j}^{\mathsf{R},r} = \left(\boldsymbol{z}_{i,j}^r\right)^\top \boldsymbol{b}_i^r$ which is a linear combination of regressors $\boldsymbol{z}_{i,j}^r$ derived from the full covariate information $\mathcal{C}_{i,j}$ with the subject-specific vector of random effects $\boldsymbol{b}_i^r$ of dimension d_r^{R}, captures the differences between units.

Given the random effects, it holds

$$\mathsf{E}\left[\boldsymbol{Y}_i^r \mid \boldsymbol{b}_i^r; \mathcal{C}_i\right] = \eta_{i,j}^r = \eta_{i,j}^{\mathsf{F},r} + \eta_{i,j}^{\mathsf{R},r} \quad \text{and} \quad \mathsf{var}\left[\boldsymbol{Y}_i^r \mid \boldsymbol{b}_i^r; \mathcal{C}_i\right] = \sigma_r^2 \mathbb{I}_{n_i}$$

due to the conditional independence. However, integrating over the latent random effects we obtain the marginal distribution of the observed outcomes of the following characteristics:

$$\begin{aligned}\mathsf{E}\left[\boldsymbol{Y}_i^r \,|\, \mathcal{C}_i\right] &= \mathsf{E}\Big(\mathsf{E}\Big[\boldsymbol{Y}_i^r \Big| \boldsymbol{b}_i^r; \mathcal{C}_i\Big]\Big) = \mathbb{X}_i^r \boldsymbol{\beta}_r + \mathbb{Z}_i^r \mathbf{0} = \mathbb{X}_i^r \boldsymbol{\beta}_r, \\ \mathsf{var}\left[\boldsymbol{Y}_i^r \,|\, \mathcal{C}_i\right] &= \mathsf{E}\Big(\mathsf{var}\Big[\boldsymbol{Y}_i^r \Big| \boldsymbol{b}_i^r; \mathcal{C}_i\Big]\Big) + \mathsf{var}\Big(\mathsf{E}\Big[\boldsymbol{Y}_i^r \Big| \boldsymbol{b}_i^r; \mathcal{C}_i\Big]\Big) = \sigma_r^2 \mathbb{I}_{n_i} + \left(\mathbb{Z}_i^r\right)^\top \boldsymbol{\Sigma}_r \mathbb{Z}_i^r,\end{aligned}$$

where

$$\mathbb{X}_i^r = \begin{pmatrix} \left(\boldsymbol{x}_{i,1}^r\right)^\top \\ \vdots \\ \left(\boldsymbol{x}_{i,n_i}^r\right)^\top \end{pmatrix} \quad \text{and} \quad \mathbb{Z}_i^r = \begin{pmatrix} \left(\boldsymbol{z}_{i,1}^r\right)^\top \\ \vdots \\ \left(\boldsymbol{z}_{i,n_i}^r\right)^\top \end{pmatrix}$$

are the model regression matrices for unit i. Generally, the outcomes $Y_{i,j}^r$ become dependent across all $j = 1, \dots, n_i$. In the simplest case of simple random intercept model $\mathbb{Z}_i^r = \mathbf{1}_{n_i}$ and, hence, $\mathsf{cov}\left[Y_{i,j_1}^r, Y_{i,j_2}^r \,\Big|\, \mathcal{C}_i\right] = \boldsymbol{\Sigma}_r + \mathbb{1}_{(j_1 = j_2)}\sigma_r^2$ for $j_1, j_2 = 1, \dots, n_i$.

The classical normal linear mixed-effects model is viable only for the numeric outcomes $\mathcal{R}^{\mathsf{Num}}$, where the assumption of normality is plausible. However, for the other types of outcomes (count, binary, ordinal, general categorical) we have to adequately adapt the model. There are however, two main paths we could take. The first introduced in Section 2.1 transfers the binary or ordinal case to the classical normal LME case by cutting latent numeric outcomes. In Section 2.2, we present a more general approach through *generalized linear mixed-effects models* (GLMM) which cover even more types of outcomes. For these two sections, the specification of the distribution for the underlying random effects will be silently avoided, only to be specified for all outcomes in the following Section 2.3 with the aim to cover possible associations among the outcomes.

2.1 Threshold concept

McCullagh (1980) provides a discussion for regression models for ordinal types of outcomes. The key to modelling such outcomes is the parametrization of the probabilities of reaching each of the values. Since many ordinal outcomes arise from numeric outcomes by grouping values in given intervals, it is only natural to utilize this very idea for any observed ordinal outcome. Bock and Lieberman (1970) used this very idea for modelling dichotomous items, which is also termed as *latent variable model for dichotomous variables* (Long, 1997). Bock (1972) also used the *threshold concept* for estimating the item parameters and latent abilities when responses are scored in two or more nominal categories.

Assume that for each ordinal outcome category $Y_{i,j}^r \in \{0, \dots, K^r - 1\}$ there exists a latent numeric outcome $Y_{i,j}^{\star,r}$ responsible for the observed category. This latent outcome is purely artificial and should be viewed as an ability to excel (high values) or to fail (low values). From the data generating point of view, we will assume that, first, the latent numeric outcome value has been generated and then the category was determined based on the segmentation into intervals given by a set of thresholds $-\infty = \gamma_{-1}^r < \gamma_0^r < \cdots < \gamma_{K^r-1}^r = \infty$:

$$Y_{i,j}^r = k \quad \Longleftrightarrow \quad \gamma_{k-1}^r < Y_{i,j}^{\star,r} \leq \gamma_k^r, \tag{2.4}$$

which is called the *threshold concept* (see, e.g. Albert and Chib, 1993).

In practise, neither the latent outcomes (and their distribution) nor the thresholds are known. Since we are not limited in the underlying distribution, we can choose the normal distribution for its nice properties. In our world of longitudinal data, we suppose the classical normal LME model (2.1) for the latent outcomes $Y^{\star,r}_{i,j}$:

$$Y^{\star,r}_{i,j} \,\Big|\, \boldsymbol{b}^r_i; \mathcal{C}^r_{i,j} \sim \mathsf{N}\left(\eta^r_{i,j},\, 1\right). \tag{2.5}$$

Note that we have fixed the variance of the error terms to 1 for identifiability purposes. For the same reason (identifiability of location), one of the thresholds has to be also fixed, e.g. the first one $\gamma^r_0 = 0$, otherwise with a shift in the distribution one could shift the thresholds correspondingly. We will denote by $\boldsymbol{\gamma}^r = \left(\gamma^r_1, \ldots, \gamma^r_{K^r-2}\right)^\top$ the unknown ordered thresholds. Hedeker (2008) provides a discussion for equivalence of threshold models with GLMM analogies. With normally distributed latent variables we end up with the *probit regression model.*

In the presence of latent variables, it is natural to use a version of the EM-algorithm (Dempster et al., 1977) for finding the corresponding maximum likelihood estimators of the unknown parameters. However, the E-step requires non-trivial approximations of integrals. Hence, in this thesis we focus on a Bayesian solution to this problem which simply considers latent variables as additional model parameters and avoids evaluation of the difficult integrals. Due to hierarchical structure of our model it is straightforward to derive the full-conditional distributions which are used to construct a Gibbs sampler, see Chapter 5 for more details.

Binary outcomes could be viewed as a special type of ordinal outcomes, because the underlying numeric outcomes could be interpreted as a propensity for success. In such a case of $K^r = 2$, there are no unknown thresholds to be estimated. Thresholding of a binary outcome and an ordinal outcome from the `PBC910` dataset is depicted in Figure 2.1, where the density of normal distribution behind is illustratively selected to fit the probability proportions to the

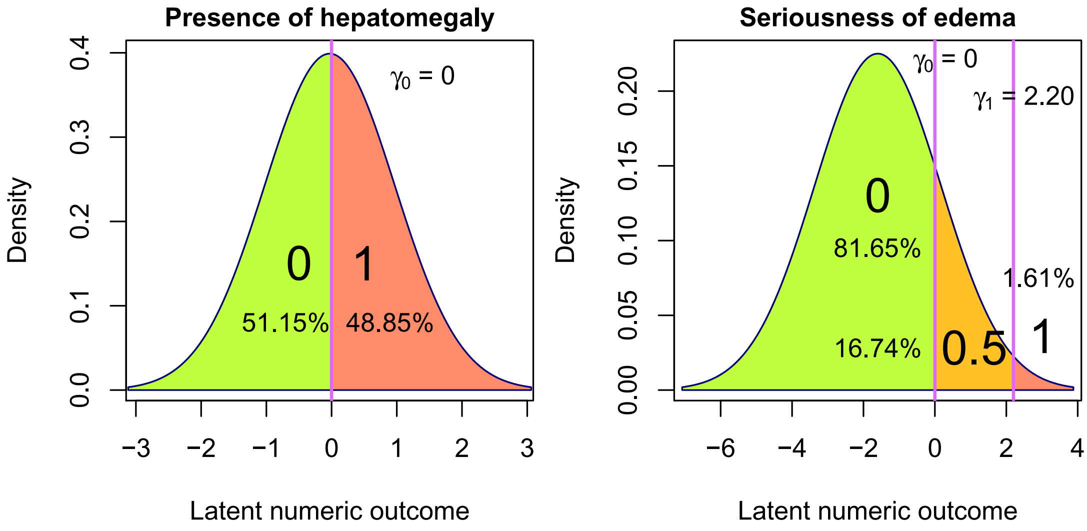


Figure 2.1: `PBC910` dataset. Latent thresholding of a binary (left) and an ordinal outcome (right).

estimated thresholds ($\gamma_0 = 0$, for γ_1 see Table 2.1).

The binary and ordinal outcomes modelled by this latent variable thresholding approach will be denoted by $\mathsf{t}(r) = \mathsf{OB}$. Given random effects, a single observation has the following pdf:

$$p_{\mathsf{OB}}\left(Y^r_{i,j} = y \,\middle|\, \boldsymbol{b}^r_i, \boldsymbol{\gamma}^r, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) = \int\limits_{\gamma^r_{y-1}}^{\gamma^r_y} p_{\mathsf{N}}\left(Y^{\star,r}_{i,j} = y^\star \,\middle|\, \boldsymbol{b}^r_i, \tau_r = 1, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) \mathrm{d}y^\star, \tag{2.6}$$

where the latent numeric outcome behind has to be integrated out on the corresponding interval, bounds of which are given by the thresholds and the observed level $y \in \{0, \ldots, K^r - 1\}$.

Bock (1972) provides a methodology for modelling any general categorical outcome, where the lack of ordering would contradict the ordinality of a single latent numeric outcome behind in emphthreshold concept. There has to be assumed more than one underlying latent numeric outcome to circumvent this problem. To be precise, an *extremal concept* is taken where the category y is associated with the underlying latent *latent tendency* for that category which is maximal. Hedeker (2008) points out that it is equivalent to multinomial regression, which will be introduced later in Section 2.2.4.

2.2 Generalized linear mixed-effects models

Generalized linear mixed-effects models (GLMM) extend the classical LME model to any distributional family of exponential type, where the conditional expected value $\mathsf{E}\left[\boldsymbol{Y}^r_i \,\middle|\, \boldsymbol{b}_i; \mathcal{C}^r_i\right]$ is tied with the linear predictor $\eta^r_{i,j}$ through an appropriate link function. Proper definition of GLMM can be found in Jiang (2007). Here we will satisfy with specific model choices for our needs.

This extension to GLMM opens up several potential models for numeric variables by different combination of distributional families (normal, Gamma, inverse Gaussian, . . .) and link functions (identity, logarithm, inverse). Nevertheless, the classical LME is sufficient in our real data applications, hence, we rather focus on models for different data types. Except for the classical LME, we will use log-linear mixed model for count outcomes $\mathcal{R}^{\mathsf{Poi}}$, logistic regression with random effects for binary outcomes $\mathcal{R}^{\mathsf{Bin}}$, ordinal logit regression for ordinal outcomes $\mathcal{R}^{\mathsf{Ord}}$ and, finally, multinomial logit regression for general categorical outcomes $\mathcal{R}^{\mathsf{Cat}}$. Our methodology could be extended for any combination of distributional family and link function, however, for the sake of simplicity and clarity only these five different types of models will be considered.

2.2.1 Model for count outcomes

Count variables ($r \in \mathcal{R}^{\mathsf{Poi}}$) take values in $\mathbb{N}_0$. Poisson distribution is particularly useful when modelling count variables expressing the total number of events that has occurred during a certain time period. Negative-binomial distribution would be useful when counting number of trials before k-th success. And one could find even more examples of distributions for count variables depending on the interpretation.

Here we will assume only the Poisson count data, for which the corresponding GLMM with canonical log link is called *log-linear* mixed-effects model. Since $Y_{i,j}^r \,\middle|\, \boldsymbol{b}_i^r \sim \mathsf{Pois}\left(\exp\{\eta_{i,j}^r\}\right)$, we also have $\mathsf{E}\left[Y_{i,j}^r \,\middle|\, \boldsymbol{b}_i^r\right] = \mathsf{var}\left[Y_{i,j}^r \,\middle|\, \boldsymbol{b}_i^r\right] = \exp\{\eta_{i,j}^r\}$. Given random effects, a single Poisson count observation $Y_{i,j}^r \in \mathbb{N}_0$ has the pdf

$$p_{\mathsf{P}}\left(Y_{i,j}^r = y \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) \propto \exp\left\{y\eta_{i,j}^r - \exp\{\eta_{i,j}^r\}\right\} \tag{2.7}$$

and contributes to the log-likelihood by

$$\ell_{\mathsf{P}}\left(Y_{i,j}^r = y \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) = \text{const.} + y\eta_{i,j}^r - \exp\{\eta_{i,j}^r\}. \tag{2.8}$$

If the number $Y_{i,j}^r$ can be interpreted as a number of events during $o_{i,j}^r$ units of time, then the distribution should rather be $Y_{i,j}^r \,\middle|\, \boldsymbol{b}_i^r \sim \mathsf{Pois}\left(o_{i,j}^r \exp\{\eta_{i,j}^r\}\right)$. In such a case, we have to redefine the predictor $\eta_{i,j}^r$ to include an *offset*, a covariate of fixed effect size of 1. In particular, $\eta_{i,j}^r := \eta_{i,j}^{\mathsf{O},r} + \eta_{i,j}^{\mathsf{F},r} + \eta_{i,j}^{\mathsf{R},r}$, where $\eta_{i,j}^{\mathsf{O},r}$ is the offset part which here takes the form of $\eta_{i,j}^{\mathsf{O},r} = \log o_{i,j}^r$.

2.2.2 Model for binary outcomes

Binary outcomes $r \in \mathcal{R}^{\mathsf{Bin}}$ are assumed to follow Bernoulli trial distribution. The probability of success is linked to the predictor by a suitable link function. One can choose from wide variety of link functions. For example, the *probit* link (quantile function of standard normal distribution) would correspond to modelling by latent normally distributed variable by the threshold concept. Hence, here we will avoid that option and focus on the canonical *logit* link $\mathsf{logit}\, p = \log \frac{p}{1-p}$.

The probability of a success in logistic mixed-effects regression is linked to the linear predictor by the inverse logit function:

$$\mathsf{P}\left[Y_{i,j}^r = 1 \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right] = \mathsf{logit}^{-1}\left(\eta_{i,j}^r\right) = \frac{\exp\left\{\eta_{i,j}^r\right\}}{1 + \exp\left\{\eta_{i,j}^r\right\}}.$$

Given random effects, a single binary observation $y = Y_{i,j}^r \in \{0, 1\}$ has the pdf

$$p_{\mathsf{B}}\left(Y_{i,j}^r = y \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) = \left[\mathsf{logit}^{-1}\left(\eta_{i,j}^r\right)\right]^y \left[1 - \mathsf{logit}^{-1}\left(\eta_{i,j}^r\right)\right]^{1-y}, \tag{2.9}$$

and contributes to the log-likelihood by

$$\ell_{\mathsf{B}}\left(Y_{i,j}^r = y \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) = y\eta_{i,j}^r - \log\left(1 + \exp\left\{\eta_{i,j}^r\right\}\right). \tag{2.10}$$

2.2.3 Model for ordinal outcomes

There are several ways, how to generalize logit probability parametrization for the case of ordinal outcomes $r \in \mathcal{R}^{\mathsf{Ord}}$. We will model the cumulative probabilities by cumulative logits. Hedeker (2008, Sec 6.3) points out that it is equivalent to the threshold concept applied to latent variables following a logistic distribution. An-alternative would be to use adjacent-categories logit (Hartzel et al., 2001, Sec. 2.2), where logarithm of ratio of probabilities of adjacent categories is modelled by predictor shifted by category-specific intercept.

Let us have an ordinal outcome of K^r levels. The cumulative probabilities are modelled by

$$p_k := \mathsf{P}\left[Y_{i,j}^r > k \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r, \boldsymbol{c}_r; \mathcal{C}_{i,j}\right] = \mathsf{logit}^{-1}\left(\eta_{i,j}^r - c_{r,k}\right) \tag{2.11}$$

for $k \in \{0, 1, \dots, K^r - 1\}$, where $-\infty = c_{r,-1} < c_{r,0} < c_{r,1} < \cdots < c_{r,K^r-1} = \infty$ are unknown ordered intercepts which shift the predictor free of an intercept term within the fixed part for identifiability purposes. We denote by $\boldsymbol{c}_r = (c_{r,0}, c_{r,1}, \dots, c_{r,K^r-2})^\top$ the unknown values. The probabilities p_k are decreasing with k, we set $p_{-1} = \mathsf{P}\left[Y_{i,j}^r > -1 \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r, \boldsymbol{c}_r; \mathcal{C}_{i,j}\right] = 1$ and end with zero probability $p_{K^r-1} = \mathsf{P}\left[Y_{i,j}^r > K^r - 1 \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r, \boldsymbol{c}_r; \mathcal{C}_{i,j}\right] = 0$. Given random effects, a single ordinal observation $y = Y_{i,j}^r \in \{0, 1, \dots, K^r - 1\}$ has the pdf given by a difference of adjacent cumulative probabilities

$$p_{\mathsf{O}}\left(Y_{i,j}^r = y \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r, \boldsymbol{c}_r; \mathcal{C}_{i,j}\right) = q_y := p_{y-1} - p_y \tag{2.12}$$

and contributes to the log-likelihood by

$$\ell_{\mathsf{O}}\left(Y_{i,j}^r = y \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) = \log q_y. \tag{2.13}$$

Under $K^r = 2$ this model specification would reduce to the previous case of logistic regression model since the only finite term $c_{r,0}$ would play the role of intercept term within the predictor.

Note that this model formulation is based on the *proportional odds* assumption; the log-odds differ only in the intercepts: $\log(p_k/(1-p_k)) = \eta_{i,j}^r - c_{r,k}, k = 0, \dots, K^r - 2$. In case this assumption is violated, we would have to assume different $\boldsymbol{\beta}_r$ parameters for each category level. That would result in a set of different predictors, which is used later for the raw (non-cumulative) probabilities when modelling general categorical outcomes. A compromise would then be to identify the covariates which violate the proportional odds assumption and make only the corresponding fixed effects category-specific. However, we will avoid this *partial proportional odds model* introduced by Peterson and Harrell (1990) for simplicity of the model formulation since we later want to create a mixture of the specified models.

2.2.4 Model for general categorical outcomes

Here we work with general categorical (nominal) outcomes $r \in \mathcal{R}^{\mathsf{Cat}}$ which lack any natural ordering. Cumulative probabilities now do not have any meaning since the levels could be arbitrarily permuted without any consequence. Hence, instead of generalizing the previous model for ordinal outcomes we will make each of the probabilities proportional to the exp of category-specific predictor. In literature this model is called multinomial logistic regression model with random effects (Hedeker, 2008, Sec. 6.4).

Let us have a nominal outcome of $K^r > 2$ unordered levels. Each category level $k \in \{0, 1, \dots, K^r - 1\}$ is supposed to have its own linear predictor $\eta_{i,j,k}^r$, collectively we denote $\boldsymbol{\eta}_{i,j}^r = \left(\eta_{i,j,0}^r, \dots, \eta_{i,j,K^r-1}^r\right)$. It is formed by $\eta_{i,j,k}^{r,\mathsf{F}} = \left(\boldsymbol{x}_{i,j}^r\right)^\top \boldsymbol{\beta}_{r,k}$ and $\eta_{i,j,k}^{r,\mathsf{r}} = \left(\boldsymbol{x}_{i,j}^r\right)^\top \boldsymbol{b}_{i,k}^r$ where we have category-specific fixed effects $\boldsymbol{\beta}_{r,k}$ and random effects $\boldsymbol{b}_{i,k}^r$. When $r \in \mathcal{R}^{\mathsf{Cat}}$ and $\boldsymbol{\beta}_r$ or $\boldsymbol{b}_i^r$ miss the index k, we understand by this notation a collection of these effects across all the possible k. The probability of attaining level $k \in \{0, 1, \dots, K^r - 1\}$ is then proportional to

$$\mathsf{P}\left[Y^r_{i,j} = k \,\middle|\, \boldsymbol{b}^r_i, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right] \propto \exp\{\eta^r_{i,j,k}\}.$$

For identifiability purposes we have to fix $\eta^r_{i,j,0} = 0$ by $\boldsymbol{\beta}_{r,0} = \mathbf{0}$ and $\boldsymbol{b}_{i,0} = \mathbf{0}$. The effects are then interpreted as comparison of log odds compared to the zero level category since $\log\left(\mathsf{P}\left[Y^r_{i,j} = k_1 \,\middle|\, \cdots\right] \middle/ \mathsf{P}\left[Y^r_{i,j} = k_2 \,\middle|\, \cdots\right]\right) = \eta^r_{i,j,k_1} - \eta^r_{i,j,k_2}$. Bock (1972) even suggests to allow for any possible set of $K^r - 1$ contrasts, which we avoid for simplicity. Another way to simplify the model would be to fix the random effects to a single set of random effects $\boldsymbol{b}^r_i = \boldsymbol{b}^r_{i,1} = \cdots = \boldsymbol{b}^r_{i,K^r-1}$, which can substantially reduce the dimension of random effects. In such a case, the random effects compare the zero category with any other category, hence, the zero category should correspond to some reasonable baseline.

Finally, we can write down the resulting formulas for pdf and log-likelihood, where we use the multivariate vector *softmax* function. Given random effects, a single general categorical observation $y = Y^r_{i,j} \in \{0, 1, \dots, K^r - 1\}$ has the pdf

$$p_{\mathsf{C}}\left(Y^r_{i,j} = y \,\middle|\, \boldsymbol{b}^r_i, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) = \mathsf{softmax}_y(\boldsymbol{\eta}^r_{i,j}) = \frac{\exp\{\eta^r_{i,j,y}\}}{1 + \sum_{k=1}^{K^r-1} \exp\{\eta^r_{i,j,k}\}} \tag{2.14}$$

and contributes to the log-likelihood by

$$\ell_{\mathsf{C}}\left(Y^r_{i,j} = y \,\middle|\, \boldsymbol{b}^r_i, \boldsymbol{\beta}_r; \mathcal{C}_{i,j}\right) = \eta^r_{i,j,y} - \log\left(1 + \sum_{k=1}^{K^r-1} \exp\{\eta^r_{i,j,k}\}\right). \tag{2.15}$$

Note that in case of $K^r = 2$ there would be only a single unknown predictor related to $y = 1$, which gives exactly the logistic regression model presented above in Section 2.2.2.

2.3 Random effects distribution

So far the models for outcomes were treated as independent of each other. However, that could hardly be used in practice since the outcomes in real datasets are often highly correlated. We will overcome this problem by a simple trick with random effects.

In the framework of mixed-effects models, the random effects $\boldsymbol{b}^r_i$ capture the correlation between the outcome values observed for each unit i and outcome $r \in \mathcal{R}$ conditional on the regression model. In the multivariate setting with several different outcome variables, the random effects are also used to capture correlations between different outcome variables for a unit i. To this end, we suppose a joint multivariate distribution for all random effects inspired by the work of Fieuws and Verbeke (2004, 2006) who explore the bivariate case in detail. Many researchers, including Komárek and Komárková (2013), have used this methodology to join mixed models for responses of different type.

Let us denote the vector of random effects for subject i by $\boldsymbol{b}_i = \{\boldsymbol{b}^r_i, r \in \mathcal{R}\}$. The overall random effects vector $\boldsymbol{b}_i$ is now assumed to follow a centred multivariate normal distribution with a *general* covariance matrix, i.e. it is assumed

$$\boldsymbol{b}_i \overset{\text{iid}}{\sim} \mathsf{N}_{d^{\mathsf{R}}}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2.16}$$

where $d^{\mathsf{R}} = \sum_{r \in \mathcal{R}} d^{\mathsf{R}}_r$ is the total dimension of $\boldsymbol{b}_i$ and $\boldsymbol{\Sigma} > 0$ is the positive-definite covariance matrix of the random effects. A general structure is assumed for this matrix thus allowing to capture arbitrary within-subject dependencies between the different outcomes. However, the key assumption here is that the outcomes are independent given the random effects. This independence disappears when transitioning to marginal distribution of the outcomes.

Consider, for example, the case $|\mathcal{R}^{\mathsf{Num}}| = |\mathcal{R}| = 2$. The two numeric outcomes are independent given random effects. Assume simple random intercept term, i.e. $\eta^{r,\mathsf{R}}_{i,j} = b^r_i$. Then,

$$\mathsf{var}\begin{pmatrix} Y^1_{i,j_1} \\ Y^2_{i,j_2} \end{pmatrix} = \mathsf{E}\left(\mathsf{var}\left[\begin{pmatrix} Y^1_{i,j_1} \\ Y^2_{i,j_2} \end{pmatrix} \middle| \boldsymbol{b}_i; \mathcal{C}_i\right]\right) + \mathsf{var}\left(\mathsf{E}\left[\begin{pmatrix} Y^1_{i,j_1} \\ Y^2_{i,j_2} \end{pmatrix} \middle| \boldsymbol{b}_i; \mathcal{C}_i\right]\right) =$$

$$= \begin{pmatrix} \sigma^2_1 & 0 \\ 0 & \sigma^2_2 \end{pmatrix} + \mathsf{var}\begin{pmatrix} \eta^{1,\mathsf{R}}_{i,j_1} \\ \eta^{2,\mathsf{R}}_{i,j_2} \end{pmatrix} = \begin{pmatrix} \sigma^2_1 & 0 \\ 0 & \sigma^2_2 \end{pmatrix} + \mathsf{var}\left(\boldsymbol{b}_i\right) = \begin{pmatrix} \sigma^2_1 & 0 \\ 0 & \sigma^2_2 \end{pmatrix} + \boldsymbol{\Sigma}.$$

In this particular setting the non-diagonal elements of $\boldsymbol{\Sigma}$ correspond to covariances between observations of different outcomes. However, how exactly the variance matrix $\boldsymbol{\Sigma}$ effects the relationships between outcomes of different type is difficult to express.

Figure 2.2 demonstrates how the value of a correlation coefficient ρ between random intercepts of simulated numeric and binary longitudinal outcomes affects the marginal dependencies. As expected, positive correlation increases the odds with numeric outcome and vice versa for the negative correlation, while zero correlation yields no marginal relationship between the two outcomes.

Indisputably, the model is able to capture some marginal relationships and incorporates them at least partially through the variance matrix $\boldsymbol{\Sigma}$, although the associations could probably be more precisely captured by, e.g. copulas (Nelsen, 1999). Simultaneously, it allows us to work with the individual observations as independent given the random effects and covariates, which will be heavily exploited in the estimation part.

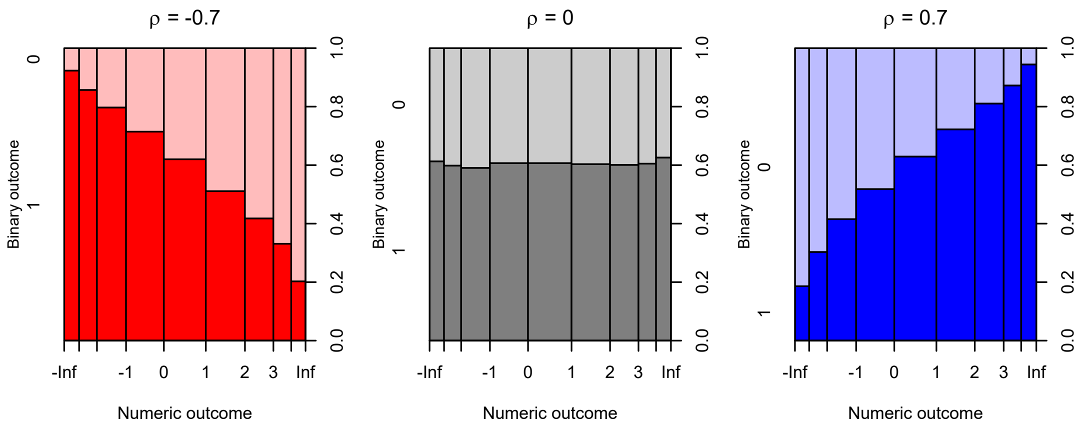


Figure 2.2: Ratios of binary outcome values across different factorized values of the numeric outcome of the simulated longitudinal dataset for $n = 10\,000$ subjects each of $n_i = 4$ observations connected through random intercepts with correlation $\rho \in \{-0.7, 0, 0.7\}$.

2.4 Models for longitudinal outcomes of a mixed type

We have proposed several approaches for modelling the individual outcomes. Would we have to consider all possible combinations of all approaches including the suggested alternatives we would get an immeasurable number of different statistical models. Hence, we dedicate this section to introduce the two main considered approaches. The initial model (Vávra and Komárek, 2022) considers only numeric outcomes and the threshold concept for modelling categorical outcomes. The second more advanced model (Vávra et al.) consists of individual GLMMs.

For both options, we combine all the models for individual outcomes to evaluate what is the overall contribution of one single unit to the (log-)likelihood which would then consist of n such independent blocks. Generally, such contribution of the observed outcomes $\mathbb{Y}_i$ can be expressed as

$$p\left(\mathbb{Y}_i|\boldsymbol{\beta},\ldots,\boldsymbol{\Sigma};\mathcal{C}_i\right) = \int \prod_{r\in\mathcal{R}}\prod_{j=1}^{n_i} p_{\mathsf{t}(r)}\left(Y_{i,j}^r \,\middle|\, \boldsymbol{b}_i^r,\boldsymbol{\beta}_r,\ldots;\mathcal{C}_i\right)\cdot p(\boldsymbol{b}_i|\boldsymbol{\Sigma})\,\mathrm{d}\boldsymbol{b}_i, \qquad (2.17)$$

where $\boldsymbol{\beta} = \{\boldsymbol{\beta}_r, r\in\mathcal{R}\}$, the dots stand for other unknown model parameters and the unobserved random effects $\boldsymbol{b}_i$ have to be integrated out.

2.4.1 The threshold concept model

The first model we proposed (Vávra and Komárek, 2022) combined the classical normal LME for numeric outcomes $\mathcal{R}^{\mathsf{Num}}$ and the threshold concept for binary and ordinal outcomes collectively denoted by $\mathcal{R}^{\mathsf{OB}} = \mathcal{R}^{\mathsf{Bin}}\cup\mathcal{R}^{\mathsf{Ord}}$ and $\mathcal{R} = \mathcal{R}^{\mathsf{Num}}\cup\mathcal{R}^{\mathsf{OB}}$ since count or nominal outcomes were not considered. This model will be from now on referred to as the *threshold concept* model.

Let us have a unit i and denote by $\mathbb{Y}_i^{\mathsf{N}}$ and $\mathbb{Y}_i^{\mathsf{OB}}$ all the outcome observations of the corresponding type; $\mathbb{Y}_i^{\mathsf{N}} = \{Y_{i,j}^r, j=1,\ldots,n_i, r\in\mathcal{R}^{\mathsf{Num}}\}$, $\mathbb{Y}_i^{\mathsf{OB}} = \{Y_{i,j}^r, j = 1,\ldots,n_i, r\in\mathcal{R}^{\mathsf{OB}}\}$. For the categorical outcomes we analogously define the corresponding set of all latent numeric outcomes by $\mathbb{Y}_i^{\star,\mathsf{OB}}$. One of the benefits of this model is that the collection $\mathbb{Y}_i^{\star} = \{\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}}\}$ follows multivariate normal distribution. Note that $Y_{i,j}^r = Y_{i,j}^{\star,r}$ for $r\in\mathcal{R}^{\mathsf{Num}}$ in this notation.

First, we have to identify the latent unobserved elements – the random effects $\boldsymbol{b}_i$ and latent numeric outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$. The pdf for the observed data $\mathbb{Y}_i = \{\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\mathsf{OB}}\}$ is then obtained by integrating the latent elements out of the joint pdf for $\{\mathbb{Y}_i, \boldsymbol{b}_i, \mathbb{Y}_i^{\star,\mathsf{OB}}\}$. After a slight regrouping of factors in (2.17) we obtain

$$\begin{aligned} &p\left(\mathbb{Y}_i|\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\gamma},\boldsymbol{\Sigma};\mathcal{C}_i\right) = \\ &= \int\!\!\int \underbrace{p\left(\mathbb{Y}_i^{\mathsf{OB}}\,\middle|\,\mathbb{Y}_i^{\star,\mathsf{OB}},\boldsymbol{\gamma}\right)}_{\text{threshold concept (2.4)}}\cdot\underbrace{p\left(\mathbb{Y}_i^{\mathsf{N}},\mathbb{Y}_i^{\star,\mathsf{OB}}\,\middle|\,\boldsymbol{b}_i,\boldsymbol{\beta},\boldsymbol{\tau};\mathcal{C}_i\right)}_{\text{MV LME (2.1),(2.5)}}\cdot\underbrace{p\left(\boldsymbol{b}_i|\boldsymbol{\Sigma}\right)}_{(2.16)}\,\mathrm{d}\boldsymbol{b}_i\,\mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}, \end{aligned} \qquad (2.18)$$

where $\boldsymbol{\beta} = \{\boldsymbol{\beta}_r, r\in\mathcal{R}\}$, $\boldsymbol{\tau} = \{\boldsymbol{\tau}_r, r\in\mathcal{R}^{\mathsf{Num}}\}$, $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_r, r\in\mathcal{R}^{\mathsf{Ord}}\}$. The first factor only declares the bounds for integration chosen from $\boldsymbol{\gamma}$ according to the observed categories $\mathbb{Y}_i^{\mathsf{OB}}$:

$$p\left(\mathbb{Y}_i^{\mathsf{OB}} \,\middle|\, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{\gamma}\right) = \prod_{r \in \mathcal{R}^{\mathsf{OB}}} \prod_{j=1}^{n_i} \mathbb{1}_{\left(\gamma^r_{y^r_{i,j}-1}, \gamma^r_{y^r_{i,j}}\right]} \left(y_{i,j}^{\star,r}\right). \tag{2.19}$$

The second factor is dedicated to numeric and latent numeric outcomes, for which the classical normal LME is supposed, hence, given the random effects, there appears a product of pdfs of univariate normal distribution:

$$p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \,\middle|\, \boldsymbol{b}_i, \boldsymbol{\beta}, \boldsymbol{\tau}; \mathcal{C}_i\right) = \prod_{r \in \mathcal{R}^{\mathsf{Num}}} \prod_{j=1}^{n_i} \varphi\left(y_{i,j}^r; \eta_{i,j}^r, \tau_r^{-1}\right) \prod_{r \in \mathcal{R}^{\mathsf{OB}}} \prod_{j=1}^{n_i} \varphi\left(y_{i,j}^{\star,r}; \eta_{i,j}^r, 1\right). \tag{2.20}$$

The last factor is again a multivariate normal distribution density coming from our assumption on the random effects (2.16):

$$p\left(\boldsymbol{b}_i | \boldsymbol{\Sigma}\right) = \varphi\left(\boldsymbol{b}_i; \boldsymbol{0}; \boldsymbol{\Sigma}\right) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{b}_i\right\}. \tag{2.21}$$

The double integral in (2.18) could be evaluated only when no categorical outcomes are observed, $\mathcal{R} = \mathcal{R}^{\mathsf{Num}}$ since (2.20) and (2.21) conjugate into the shape of multivariate normal density in $\boldsymbol{b}_i$, which facilitates the integration. However, in the presence of categorical outcomes the situation becomes problematic, especially, with high dimension of $\mathbb{Y}_i^{\mathsf{OB}}$ because integration of multivariate normal density over a given multidimensional interval consists of $2^{n_i|\mathcal{R}^{\mathsf{OB}}|}$ summands in general. In our estimation process via MCMC we elegantly avoid the necessity for evaluation of such integrals. However, for some applications (e.g. calculation of classification probabilities) we employ numerical methods (Algorithm 4 by Genz) for approximation of the integral, see Section 7.2.

`PBC910` analysis by the *threshold concept* model

Now we demonstrate the use of the *threshold concept* model on `PBC910` dataset. Estimation is done via our MCMC sampler (Chapter 5) with $G = 1$ clusters. The implementation of this sampler did not, however, cover methods for accounting for missing outcome values. Hence, we have to use the complete data rows only, which costs 17 data rows out of the total 918.

Since count type is unavailable in this model, we decided to model the *platelet count* as a numeric outcome along with the logarithm of *serum bilirubin*. We have considered only *presence of hepatomegaly* as a binary outcome and *seriousness of edema* as an ordinal outcome. For all four outcomes we decided to use the same structure of fixed effects:

$$\eta^F = \beta_0 + \beta_A A + \beta_M M + \beta_{A:M} AM + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3, \tag{2.22}$$

where A is the age at the entry (divided by 10), M is an indicator of male patient and S_1, S_2, S_3 are B-spline bases for time since the entry (quadratic, single inner knot in 1.25). The random effects were comprised of solely random intercept, hence, the covariance matrix $\boldsymbol{\Sigma}$ is 4-dimensional (in order: `bili`, `platelet`, `hepato`, `edema`).

Table 2.1: `PBC910` dataset. Posterior medians of the *threshold concept* model parameters including 95% equal-tailed credible intervals.

Parameter	Numeric outcomes				Binary outcome		Ordinal outcome	
	Log(bilirubin)		Platelet count		Hepatomegaly		Edema	
β_0	0.92	(0.37; 1.46)	278.56	(225.64; 331.49)	−0.09	(−1.37; 1.23)	−4.30	(−6.04; −2.75)
β_A	−0.12	(−0.23; −0.01)	−0.98	(−11.52; 9.57)	−0.04	(−0.31; 0.22)	0.44	(0.14; 0.75)
β_M	−0.08	(−1.46; 1.28)	90.01	(−61.69; 238.41)	−2.01	(−5.41; 1.41)	−1.78	(−6.07; 2.20)
$\beta_{A:M}$	0.12	(−0.13; 0.37)	−22.19	(−49.64; 5.71)	0.54	(−0.11; 1.19)	0.19	(−0.54; 0.91)
β_1	−0.16	(−0.28; −0.05)	−32.55	(−48.01; −17.14)	−0.07	(−0.57; 0.44)	−0.22	(−0.83; 0.38)
β_2	0.30	(0.13; 0.47)	−12.33	(−35.78; 10.80)	0.36	(−0.42; 1.14)	1.00	(0.06; 1.96)
β_3	0.06	(−0.17; 0.28)	−43.91	(−74.06; −13.55)	0.15	(−0.89; 1.17)	0.70	(−0.51; 1.91)
$\sigma = \tau^{-\frac{1}{2}}$	0.38	(0.36; 0.40)	51.08	(48.41; 54.00)	1		1	
γ_1	-		-		-		2.21	(1.84; 2.61)

Table 2.1 provides quantile estimates of the posterior distribution of unknown parameters specific to each modelled outcome. The resulting spline parametrizations are depicted in Figure 2.3a for numeric outcomes only. We can see slightly increasing trend for *serum bilirubin* and decreasing trend for *platelet count*. When the matrix $\boldsymbol{\Sigma}$ is decomposed into elements of standard deviations and correlations, their posterior medians take the following values

$$\mathsf{diag}\begin{pmatrix} 0.89 \\ 81.37 \\ 1.85 \\ 1.90 \end{pmatrix} \begin{pmatrix} 1.00 & -0.16 & 0.55 & 0.35 \\ -0.16 & 1.00 & -0.26 & -0.21 \\ 0.55 & -0.26 & 1.00 & 0.39 \\ 0.35 & -0.21 & 0.39 & 1.00 \end{pmatrix} \mathsf{diag}\begin{pmatrix} 0.89 \\ 81.37 \\ 1.85 \\ 1.90 \end{pmatrix}$$

which proves non-negligible correlations among the latent random intercepts. Notice the negative correlations of *platelet count* with other outcomes.

2.4.2 The GLMM-based model

Having established the *threshold concept* model we sought for model allowing for more types of outcomes. Naturally, we completely switched to GLMM framework (Vávra et al.). Within that paper we established the model for numeric, binary, ordinal and general categorical outcomes and briefly mentioned the possibility for count outcomes. Here we will work with all five types of outcomes. From now on, we will call this model *GLMM-based* model.

Models from Section 2.2 are united through the joint random effects distribution. To establish a certain structure that would be kept throughout the whole thesis we divide the vector of all random effects $\boldsymbol{b}_i$ of unit i into subvectors depending on the type of outcomes they belong to. In particular, $\boldsymbol{b}_i^{\mathsf{N}} = \left\{\boldsymbol{b}_i^r, r \in \mathcal{R}^{\mathsf{Num}}\right\}$, $\boldsymbol{b}_i^{\mathsf{P}} = \left\{\boldsymbol{b}_i^r, r \in \mathcal{R}^{\mathsf{Poi}}\right\}$, $\boldsymbol{b}_i^{\mathsf{B}} = \left\{\boldsymbol{b}_i^r, r \in \mathcal{R}^{\mathsf{Bin}}\right\}$, $\boldsymbol{b}_i^{\mathsf{O}} = \left\{\boldsymbol{b}_i^r, r \in \mathcal{R}^{\mathsf{Ord}}\right\}$ and $\boldsymbol{b}_i^{\mathsf{C}} = \left\{\boldsymbol{b}_i^r, r \in \mathcal{R}^{\mathsf{Cat}}\right\}$. We sort these subvectors within $\boldsymbol{b}_i$ in the same order they have been presented above. Moreover, $\boldsymbol{b}_i$ is assumed to follow a centred multivariate normal distribution with a general covariance matrix $\boldsymbol{\Sigma}$ which will also be of block structure:

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_i^{\mathsf{N}} \\ \boldsymbol{b}_i^{\mathsf{P}} \\ \boldsymbol{b}_i^{\mathsf{B}} \\ \boldsymbol{b}_i^{\mathsf{O}} \\ \boldsymbol{b}_i^{\mathsf{C}} \end{pmatrix} \overset{\text{iid}}{\sim} \mathsf{N}_{d^{\mathsf{R}}} \left(\boldsymbol{0},\, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathsf{NN}} & \boldsymbol{\Sigma}_{\mathsf{NP}} & \boldsymbol{\Sigma}_{\mathsf{NP}} & \boldsymbol{\Sigma}_{\mathsf{NO}} & \boldsymbol{\Sigma}_{\mathsf{NC}} \\ \boldsymbol{\Sigma}_{\mathsf{PN}} & \boldsymbol{\Sigma}_{\mathsf{PP}} & \boldsymbol{\Sigma}_{\mathsf{PB}} & \boldsymbol{\Sigma}_{\mathsf{PO}} & \boldsymbol{\Sigma}_{\mathsf{PC}} \\ \boldsymbol{\Sigma}_{\mathsf{BN}} & \boldsymbol{\Sigma}_{\mathsf{BP}} & \boldsymbol{\Sigma}_{\mathsf{BB}} & \boldsymbol{\Sigma}_{\mathsf{BO}} & \boldsymbol{\Sigma}_{\mathsf{BC}} \\ \boldsymbol{\Sigma}_{\mathsf{ON}} & \boldsymbol{\Sigma}_{\mathsf{OP}} & \boldsymbol{\Sigma}_{\mathsf{OB}} & \boldsymbol{\Sigma}_{\mathsf{OO}} & \boldsymbol{\Sigma}_{\mathsf{OC}} \\ \boldsymbol{\Sigma}_{\mathsf{CN}} & \boldsymbol{\Sigma}_{\mathsf{CP}} & \boldsymbol{\Sigma}_{\mathsf{CB}} & \boldsymbol{\Sigma}_{\mathsf{CO}} & \boldsymbol{\Sigma}_{\mathsf{CC}} \end{pmatrix} \right). \tag{2.23}$$

The equation (2.17) with the use of (2.2), (2.7), (2.9), (2.12), (2.14) (pdfs for individual GLMM types) and (2.21) (random effects distribution) is actually the most space-efficient way to express the pdf for observations of a single unit i. We combine them into

$$\begin{aligned} p\left(\mathbb{Y}_i \middle| \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{c}, \boldsymbol{\Sigma}; \mathcal{C}_i\right) \propto \int \prod_{r \in \mathcal{R}^{\mathsf{Num}}} \tau_r^{\frac{n_i}{2}} \exp\left\{ -\frac{\tau_r}{2} \sum_{j=1}^{n_i} \left(y_{i,j}^r - \eta_{i,j}^r \right)^2 \right\} \cdot& \\ \cdot \prod_{r \in \mathcal{R}^{\mathsf{Poi}}} \exp\left\{ \sum_{j=1}^{n_i} \left(y_{i,j}^r \eta_{i,j}^r - \exp\{\eta_{i,j}^r\} \right) \right\} \cdot& \\ \cdot \prod_{r \in \mathcal{R}^{\mathsf{Bin}}} \prod_{j=1}^{n_i} \left[\mathsf{logit}^{-1}\left(\eta_{i,j}^r\right) \right]^{y_{i,j}^r} \left[1 - \mathsf{logit}^{-1}\left(\eta_{i,j}^r\right) \right]^{1-y_{i,j}^r} \cdot& \\ \cdot \prod_{r \in \mathcal{R}^{\mathsf{Ord}}} \prod_{j=1}^{n_i} \left[\mathsf{logit}^{-1}\left(\eta_{i,j}^r - c_{r, y_{i,j}^r - 1} \right) - \mathsf{logit}^{-1}\left(\eta_{i,j}^r - c_{r, y_{i,j}^r} \right) \right] \cdot& \\ \cdot \prod_{r \in \mathcal{R}^{\mathsf{Cat}}} \prod_{j=1}^{n_i} \frac{\exp\left\{ \eta_{i,j,y_{i,j}^r}^r \right\}}{1 + \sum\limits_{k=1}^{K^r-1} \exp\{\eta_{i,j,k}^r\}} \cdot& \\ \cdot \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{b}_i \right\} \,\mathrm{d}\boldsymbol{b}_i,& \end{aligned} \tag{2.24}$$

where $\boldsymbol{\beta} = \{\boldsymbol{\beta}_r, r \in \mathcal{R}\}$, $\boldsymbol{\tau} = \{\boldsymbol{\tau}_r, r \in \mathcal{R}^{\mathsf{Num}}\}$, $\boldsymbol{c} = \{\boldsymbol{c}_r, r \in \mathcal{R}^{\mathsf{Ord}}\}$. Equation (2.24) captures the complexity of the model we are dealing with. As elegant as the idea with random effects to jointly model the outcomes seems, it certainly complicates the evaluation of marginal distribution of the outcomes since the pieces $\boldsymbol{b}_i^r$ are scattered at different places within the predictors $\eta_{i,j}^r$. Unless there are only numeric outcomes, we cannot directly evaluate this integral. Similarly as in previous *threshold model*, this does not have to concern us that much since we aim for the Bayesian approach and the computation of integrals (2.24) will not be necessary to estimate the model by the MCMC sampling. However, for some applications (e.g. calculation of classification probabilities) we employ numerical methods (Laplacian approximation or Adaptive Gaussian Quadrature in general) for approximation of the integral, see Section 7.3.

`PBC910` analysis by the *GLMM-based* model

Again, we demonstrate the use of the just presented model on the `PBC910` dataset. Estimation is done via our MCMC sampler (Chapter 6) with $G = 1$ clusters. The implementation of this sampler is more advanced and allows for missing values among outcomes, hence, we work here with all 918 data rows.

Table 2.2: `PBC910` dataset. Posterior medians of the *GLMM-based* model parameters including 95% equal-tailed credible intervals.

Parameter	Numeric outcome		Count outcome		Binary outcome		Ordinal outcome	
	Log(bilirubin)		Platelet count		Hepatomegaly		Edema	
β_0	0.94	(0.35; 1.50)	5.56	(5.35; 5.76)	-0.15	$(-2.42; 2.09)$	-	
β_A	-0.13	$(-0.24; -0.01)$	0.00	$(-0.05; 0.04)$	-0.05	$(-0.50; 0.41)$	0.79	(0.28; 1.36)
β_M	-0.28	$(-2.07; 1.48)$	0.61	$(-0.06; 1.24)$	-4.22	$(-11.46; 2.79)$	-2.15	$(-4.89; 0.08)$
$\beta_{A:M}$	0.15	$(-0.17; 0.48)$	-0.13	$(-0.25; -0.01)$	1.05	$(-0.25; 2.40)$	0.71	$(-1.02; 2.55)$
β_1	-0.12	$(-0.23; -0.01)$	-0.13	$(-0.15; -0.11)$	-0.22	$(-1.12; 0.66)$	-0.37	$(-1.42; 0.68)$
β_2	0.21	(0.04; 0.37)	-0.05	$(-0.08; -0.02)$	0.63	$(-0.69; 2.00)$	1.67	(0.13; 3.21)
β_3	0.22	(0.01; 0.43)	-0.18	$(-0.21; -0.14)$	0.04	$(-1.80; 1.82)$	1.21	$(-0.73; 3.14)$
$\sigma = \tau^{-\frac{1}{2}}$	0.38	(0.36; 0.40)	-		-		-	
c_0	-		-		-		3.55	(2.78; 4.41)
c_1	-		-		-		7.46	(6.35; 8.68)

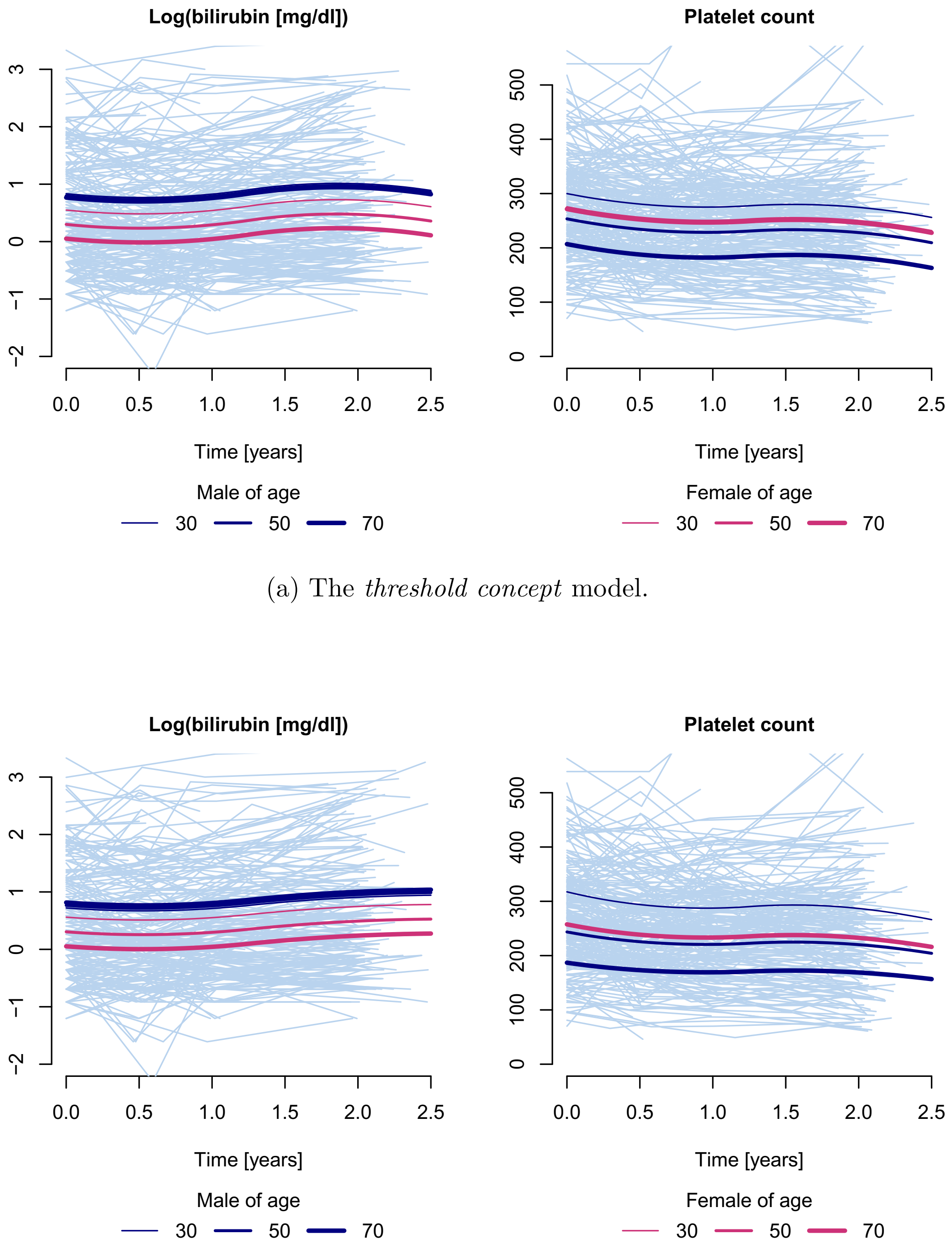
We will use the same outcomes and model structure as for the *threshold concept* model above. One small difference will be regarding the *platelet count* which can now be modelled as a count variable. Hence, we have one outcome per each type except for the general categorical case, for which there does not exist a suitable marker in `pbcseq`. The fixed effects structure will be again (2.22) with the exception of an ordinal outcome, where the intercept term β_0 is replaced by the two ordered intercepts $-\infty < c_0 < c_1 < \infty$.

We provide a table analogous to the previous one summarizing the posterior distributions by 2.5%, 50%, and 97.5% quantiles, Table 2.2. We note, however, that parameters for the non-numeric outcomes do not have identical interpretation as in the *threshold concept* model, hence, their size comparison is irrelevant maybe except for a sign (and significance). Results for the logarithm of *serum bilirubin* appear analogous. Likewise do even the estimated spline parametrizations depicted in Figure 2.3. Minor differences could be caused, for example, by those 17 additional observations. The splines for *platelet count* slightly differ due to different model behind.

Yet another parameter, interpretation of which differs from the *threshold concept* model, is the covariance matrix $\boldsymbol{\Sigma}$. We still have only random intercepts, but some of them are included in a different format. Comparing the posterior medians of standard deviations and correlations

$$\mathsf{diag}\begin{pmatrix}0.90\\0.37\\3.10\\3.23\end{pmatrix}\begin{pmatrix}1.00 & -0.17 & 0.55 & 0.34\\-0.17 & 1.00 & -0.31 & -0.27\\0.55 & -0.31 & 1.00 & 0.38\\0.34 & -0.27 & 0.38 & 1.00\end{pmatrix}\mathsf{diag}\begin{pmatrix}0.90\\0.37\\3.10\\3.23\end{pmatrix}$$

with the ones based on the previous model, we do not see any major differences in the correlation structure. On the other hand, the standard deviations for the random intercepts differ substantially.



(a) The *threshold concept* model.

(b) The *GLMM-based* model.

Figure 2.3: PBC910 dataset. Estimated spline curves for patients of different ages for males and females separately by posterior median.

3. Model-based clustering

Our primary goal is to divide observed units into groups of similar characteristics. In statistics, this task is often addressed as *unsupervised clustering* since no information about the true partition is available. Traditional methods for multivariate clustering, e.g. k-means algorithm by Hartigan and Wong (1979) (Algorithm 7), are unfit for the longitudinal nature of our mixed-type data since definition of a metric on such a sample space can be troublesome.

We will evade such metric-based methods and take a completely different approach of the *model-based clustering* (MBC), methodology introduced by Banfield and Raftery (1993) for multivariate normal distribution. We will slowly transfer from this elementary mixture of normal distributions to a general mixture of well structured probabilistic models.

3.1 Finite mixture of distributions

When the data exhibit some irregularities such as multimodality, which makes them difficult to be modelled by standard distributional families, one can create a *mixture* of distributions to better fit the data. The generating mechanism is extended by sampling the allocation indicators first (unobserved in reality) and then sampling the observed data from the distribution of the sampled group.

For simplicity, let us first assume that n independent multivariate outcomes $\boldsymbol{Y}_i, i = 1, \dots, n$ are observed. We will also assume that the heterogeneity within the data is caused by a hidden allocations of all units to one of $G > 1$ groups which differ in the distribution of $\boldsymbol{Y}_i$. The allocation will be denoted by $U_i \in \{1, \dots, G\}$ meaning $U_i = g$ if and only if unit i falls into group $g \in \{1, \dots, G\}$.

The marginal distribution of allocation indicators is fully described by parameter $\boldsymbol{w}$ which consists of probabilities

$$w_g := \mathsf{P}\left[U_i = g\right], \quad g = 1, \dots, G, \tag{3.1}$$

where $0 < w_g < 1$ and $w_1 + \dots + w_G = 1$. Then, if we denote by h_g the probability density function of $\boldsymbol{Y}_i$ given it is allocated in group g, that is, of distribution $\boldsymbol{Y}_i | U_i = g$ with pdf $h_g(\boldsymbol{y}_i) = p(\boldsymbol{Y}_i = \boldsymbol{y}_i | U_i = g)$, the marginal distribution of $\boldsymbol{Y}_i$ can be obtained from a joint distribution of $\{\boldsymbol{Y}_i, U_i\}$ by integration of latent allocation indicators, which yields us the following pdf:

$$p\left(\boldsymbol{Y}_i = \boldsymbol{y}_i\right) = \sum_{g=1}^{G} \mathsf{P}(U_i = g) p\left(\boldsymbol{Y}_i = \boldsymbol{y}_i | U_i = g\right) = \sum_{g=1}^{G} w_g h_g(\boldsymbol{y}_i).$$

One can also use Bayes' theorem to evaluate the probability of allocation unit i into group g given observed data:

$$\mathsf{P}\left[U_i = g | \boldsymbol{Y}_i = \boldsymbol{y}_i\right] = \frac{w_g h_g(\boldsymbol{y}_i)}{\sum_{l=1}^{G} w_l h_l(\boldsymbol{y}_i)} \propto w_g h_g(\boldsymbol{y}_i), \tag{3.2}$$

which is used to create the partition into clusters. However, the unknown probabilities $\boldsymbol{w}$ and model parameters $\boldsymbol{\zeta}^{(g)}$ of each h_g have to be estimated first.

The estimation by maximisation of likelihood

$$L(\boldsymbol{w}, \boldsymbol{\zeta}^{(1)}, \ldots, \boldsymbol{\zeta}^{(G)}) = \prod_{i=1}^{n} \sum_{g=1}^{G} w_g h_g(\mathbf{Y}_i; \boldsymbol{\zeta}^{(g)})$$

cannot be directly performed and has to be solved by numerical methods, e.g. by EM algorithm (Dempster et al., 1977) which simultaneously estimates the allocation probabilities. An alternative (Bayesian) approach is to assign some prior distribution to unknown parameters $\boldsymbol{w}$ and $\boldsymbol{\zeta}^{(g)}$, $g = 1, \ldots, G$ and then evaluate the posterior distribution of these unknown parameters, for more details see Chapter 4.

Let us demonstrate the mixture distribution on the well known `faithful` dataset which gathers the duration and waiting times for eruption of Old Faithful geyser in Yellowstone National Park, Wyoming, USA. Looking at Figure 3.1, the distribution appears to be bimodal (short eruption times precede short waiting times and, conversely, long eruptions times precede long waiting times). Hence, we estimated the univariate (separately for eruption and waiting time) and multivariate normal mixture model with $G = 2$ clusters and estimated the classification probabilities (3.2) as a parametric function of estimated parameters.

3.2 Model-based clustering

It did not take long time (Fraley and Raftery, 2002) to realize that the finite mixture can be used even for much more complex models. One can extend this idea to any independent blocks of data $\mathbb{Y}_i, i = 1, \ldots, n$ that follow a distribution given by the pdf h_g of a more complex form when allocated in group $g \in \{1, \ldots, G\}$. These principles are collectively referred to as *model-based clustering* (MBC) since they can be used in wide variety of statistical models.

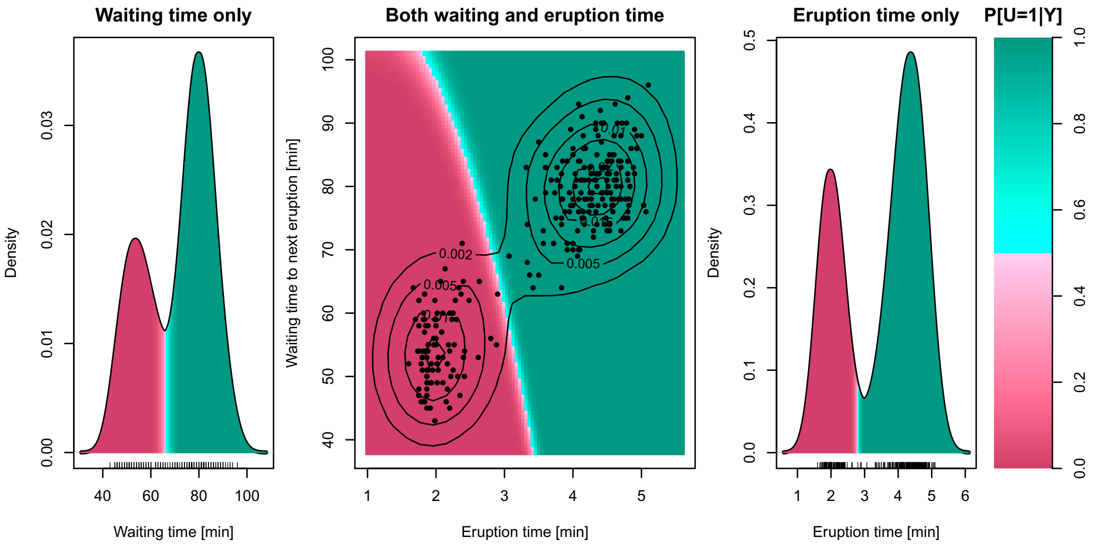


Figure 3.1: Old Faithful geyser data. Univariate (left, right) and bivariate kernel density estimates (centre). Colour depicts the probability of being allocated to one of the clusters given a data point estimated by the EM-algorithm.

Complex probabilistic models often require auxiliary variables which are not directly observed. These latent quantities will be denoted by $\mathcal{L}_i$ since they could be specific to each unit, e.g. the random effects $\boldsymbol{b}_i$ from Section 2.4. For their special role, we keep U_i separate from $\mathcal{L}_i$ in the equations within this chapter, however, in later sections we consider U_i as yet another element of $\mathcal{L}_i$ or $\mathcal{L} = \{\mathcal{L}_i, i = 1, \dots, n\}$ in general.

Let the joint distribution of observed and unobserved data $\{\mathbb{Y}_i, \mathcal{L}_i\}$ given allocation in group g depend on a set of parameters $\boldsymbol{\zeta}^{(g)}$ which can consist of several blocks of parameters (see the following section for an example). The number of these parameters could be very large and some of them would not have to be specific to each group at all. Therefore, we introduce an additional parameter $\boldsymbol{\zeta}$ which will stand for parameters common to all clusters, while $\boldsymbol{\zeta}^{(g)}$ refers to group-specific parameters only. Altogether, we denote by $\boldsymbol{\theta} = \left\{\boldsymbol{w}, \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(1)}, \dots, \boldsymbol{\zeta}^{(G)}\right\}$ the set of all unknown parameters. Additionally, the distribution of $\{\mathbb{Y}_i, \mathcal{L}_i\}$ given observed covariates $\mathcal{C}_i$ is of interest.

The pdf for $\{\mathbb{Y}_i, \mathcal{L}_i, U_i\}$ given covariates and values of unknown parameters can be decomposed into

$$p\left(\mathbb{Y}_i, \mathcal{L}_i, U_i \,|\, \boldsymbol{\theta}; \mathcal{C}_i\right) = p\left(\mathbb{Y}_i \,|\, U_i, \mathcal{L}_i, \boldsymbol{\theta}; \mathcal{C}_i\right) p\left(\mathcal{L}_i \,|\, U_i, \boldsymbol{\theta}; \mathcal{C}_i\right) p\left(U_i \,|\, \boldsymbol{\theta}\right),$$

from which the latent (unobserved) data have to be integrated out to obtain marginal distribution of $\mathbb{Y}_i$ given covariates and $\boldsymbol{\theta}$:

$$p\left(\mathbb{Y}_i \,|\, \boldsymbol{\theta}; \mathcal{C}_i\right) = \sum_{g=1}^{G} \mathsf{P}\left[U_i = g | \boldsymbol{w}\right] \int p\left(\mathbb{Y}_i, \mathcal{L}_i \,\middle|\, U_i = g, \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(g)}; \mathcal{C}_i\right) \mathrm{d}\mathcal{L}_i, \tag{3.3}$$

where $\mathrm{d}\mathcal{L}_i$ denotes integration of all unobserved elements included in $\mathcal{L}_i$ with respect to the corresponding measure. The integral in (3.3) is often intractable, which complicates the observed likelihood even more than the simple finite mixture does. The EM-algorithm for finding the MLE then has to be adjusted to deal with both allocation indicators U_i and other latent quantities $\mathcal{L}_i$. Bouveyron et al. (2019) provide several more complex examples of the use of EM-algorithm accompanied with diverse real data applications. However, we will take a Bayesian approach for model estimation since it naturally works with hierarchically well structured models, more on that later in Chapter 4.

3.2.1 Classification probabilities and rules

Analogously as in the previous section, we can use Bayes' theorem to obtain the classification probabilities given observed data and a chosen value of unknown parameters $\boldsymbol{\theta}$:

$$u_{i,g}(\boldsymbol{\theta}) := \mathsf{P}\left[U_i = g | \mathbb{Y}_i, \boldsymbol{\theta}; \mathcal{C}_i\right] \propto w_g \int p\left(\mathbb{Y}_i, \mathcal{L}_i | U_i = g, \boldsymbol{\theta}; \mathcal{C}_i\right) \mathrm{d}\mathcal{L}_i. \tag{3.4}$$

Note that one can calculate the classification probabilities even for a newly observed (or artificial) unit with the use of its set of outcomes $\mathbb{Y}_{\mathsf{new}}$ and covariates $\mathcal{C}_{\mathsf{new}}$.

To obtain the probabilities (3.4), expressions for all $g = 1, \dots, G$ have to be evaluated and then summed up to obtain the denominator. The computation for one unit and a value of $\boldsymbol{\theta}$ requires G evaluations of the integral with respect to latent data, which can become a bottle neck for the computation time, especially, when the integral has to be numerically approximated. We have dedicated several sections in Chapter 7 to explain the use of a variety of methods for integral approximations in detail.

For now, assume that we have already obtained a point estimate $\widehat{u}_{i,g}$ of (3.4), e.g. by evaluation of (3.4) at a point estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ or by a certain characteristic of the posterior distribution. The most straightforward classification of a unit i is to the cluster with the highest probability $\widehat{U}_i := \operatorname*{arg\,max}_{g\in\{1,\dots,G\}} \widehat{u}_{i,g}$. However, that may not be the best strategy for classification since there could be other competing cluster(s). Such units should rather remain unclassified ($\widehat{U}_i = 0$) by a given fixed rule. Below, we list the possible rules for classifying into $\widehat{U}_i$:

(P1) the highest probability $\widehat{u}_{i,\widehat{U}_i}$ is higher than a given limit (such as 0.5 or 0.6),

(P2) the highest probability $\widehat{u}_{i,\widehat{U}_i}$ is higher by a given margin (such as 0.2) than other classification probabilities.

If the chosen rule is not met, the unit is marked as unclassified ($\widehat{U}_i := 0$). Both rules (and many other) have their strengths and weaknesses, so one should choose carefully according to his idea of being unclassifiable.

Given interval estimates $I_g = (l_g, u_g)$ of the classification probabilities, e.g. ET or HPD intervals based on posterior distribution (see Section 4.1), we can construct even more elaborate rules for staying with $\widehat{U}_i := \operatorname*{arg\,max}_{g\in\{1,\dots,G\}} \widehat{u}_{i,g}$:

(I1) the lower bound $l_{\widehat{U}_i}$ is higher than a given threshold (such as 0.5 or 0.6),

(I2) the interval $I_{\widehat{U}_i}$ does not cross the other intervals, i.e. the lower bound $l_{\widehat{U}_i}$ is higher than any other upper bound $u_g, g \in \{1,\dots,G\} \setminus \{\widehat{U}_i\}$.

3.2.2 The number of mixture components

In the example with Old Faithful geyser at the end of Section 3.1 it was quite obvious from the distribution of the points in the scatterplot (see Figure 3.1) that we should choose $G = 2$ components. But what if the suitable number of mixture components G cannot be judged (due to data complexity) and no expert knowledge is available?

A very common solution to this problem is to simply estimate the model under different choices of $G \in \{1, 2, \dots, G_{\mathsf{max}}\}$. One then can suitably plot the resulting partition and choose the value of G leading to the visually most pleasing result. Yet, it may be difficult to judge, especially, in more complicated data structures. A more scientific approach would be to compare the fits numerically. One can compare the resulting likelihoods and whether the increase in number of unknown parameters was worth it. Akaike information criterion (AIC) and Bayesian information criterion (BIC) are favourite choices which penalize the maximal value of (log)-likelihood by the number of unknown parameters, the latter more than the former. Then one chooses the G optimizing the selected criterium across the G_{max} possible values.

For example, for the Old Faithful geyser data BIC confirms the $G = 2$ solution, while AIC prefers $G = 5$ groups. Three of the groups lie in the bottom left corner of low waiting and eruption times, the other two lie in the opposite corner. Hence, one could still interpret it as two main groups, but each requires small subdivisions to approximate the underlying distribution appropriately.

In the Bayesian setting, however, we estimate the whole posterior distribution of unknown parameters $\boldsymbol{\theta}$ instead of obtaining a point estimator $\widehat{\boldsymbol{\theta}}$ maximising the likelihood. Nevertheless, the (log)-likelihood of the model can be viewed as a parametric function of $\boldsymbol{\theta}$ and a posterior distribution of the chosen criterion can be evaluated, see Section 4.3 for the use of *deviance*. However, the Bayesian approach offers plenty of other methods for estimating the number of components. For example, we will set up the prior distribution to achieve *sparse finite mixture*, see Sections 4.2 and 6.3.

3.3 MBC for jointly modelled longitudinal outcomes of a mixed type

The challenge of modelling longitudinal data was first faced by Verbeke and Lesaffre (1996), who classified growth curves, though not explicitly called it MBC at that time. More recently, an application of similar ideas to clustering of gene-expression data is covered by Celeux et al. (2005). Subsequently, De la Cruz-Mesía et al. (2008) base their MBC procedure for longitudinal data on a non-linear mixed model. The situation of more than one continuous outcome ($|\mathcal{R}| = |\mathcal{R}^{\mathsf{Num}}| > 1$) available for clustering is considered by Villarroel et al. (2009). An early example with the use of MBC for non-continuous longitudinal data modelled by GLMM can be found in Molenberghs and Verbeke (2005, Chapter 14, Section 23.3.), but still for a single ($|\mathcal{R}| = 1$) longitudinal outcome. Proust-Lima et al. (2017) provide an R package `lcmm` for modelling several longitudinal outcomes of the same type (e.g. all binary).

The case of combining different types of longitudinal outcomes for the purpose of clustering is far more challenging and, thus, scarce in the literature. Grün and Leisch (2008) implemented a clustering algorithm (R package `flexmix`) for longitudinal data of a mixed type, however, under independence of different longitudinal outcomes measured at one occasion. This may not only be unrealistic but also prevents the analyst from exploiting information provided by the dependence structure among the outcomes in the clustering procedure. Our approach is heavily influenced by the work of Komárek and Komárková (2013, 2014) who were able to cluster longitudinal data of numeric, count and binary type and used random effects to capture the relationships among them in their R package `mixAK`. However, they lack ordinal and general categorical outcomes for the modelling of EU-SILC data where these types are common.

In the rest of this section we show how exactly the MBC methodology was used for the models introduced in Sections 2.4.1 and 2.4.2. Bayesian approach will be taken in order to estimate these two models. After the prior distributions are set in Chapter 4, the samplers are introduced in the two following Chapters 5 and 6 separately for both models. We continue in the example of the `PBC910` dataset to illustrate their setting on a real data problem.

3.3.1 MBC for the threshold concept model

Within the *threshold concept* model (see Section 2.4.1), there are several unknown parameters which can be potentially set up to be group-specific. Here, we will work under group-specificity of all parameters with the exception of $\boldsymbol{\gamma} = \{\gamma_r, r \in \mathcal{R}^{\mathsf{Ord}}\}$ which will be common to all clusters since the later derivation of the full-conditional distribution (see Section 5.1.4) would be problematic. Nevertheless, we fix some of the parameters in real-data applications. We will use the following notation: $\boldsymbol{\beta}^{(g)}$, $\boldsymbol{\tau}^{(g)}$, $\boldsymbol{\Sigma}^{(g)}$ for parameters specific to cluster $g \in \{1, \dots, G\}$. When working under MBC set up, the symbols $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{\Sigma}$ stand for collections of the group-specific parameters, that is, $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}^{(g)}, g = 1, \dots, G\right\}$, $\boldsymbol{\tau} = \left\{\boldsymbol{\tau}^{(g)}, g = 1, \dots, G\right\}$, $\boldsymbol{\Sigma} = \left\{\boldsymbol{\Sigma}^{(g)}, g = 1, \dots, G\right\}$, respectively. Altogether, in the notation from the previous section: $\boldsymbol{\zeta} = \boldsymbol{\gamma}$, $\boldsymbol{\zeta}^{(g)} = \left\{\boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{\Sigma}^{(g)}\right\}$ and $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$.

In addition, we work here with two kinds of latent variables: random effects $\boldsymbol{b}_i$ and latent numeric outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$, hence, $\mathcal{L}_i = \left\{\boldsymbol{b}_i, \mathbb{Y}_i^{\star,\mathsf{OB}}\right\}$. The pdf (3.3) for $\mathbb{Y}_i$ under a mixture of the *threshold concept* models (2.18) can be rewritten to

$$p\left(\mathbb{Y}_i \middle| \boldsymbol{\theta}; \mathcal{C}_i\right) = \sum_{g=1}^{G} w_g \int\!\!\int p\left(\mathbb{Y}_i^{\mathsf{OB}} \,\middle|\, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{\gamma}\right) \cdot \\ \cdot\, p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \,\middle|\, \boldsymbol{b}_i, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}; \mathcal{C}_i\right) \cdot p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\Sigma}^{(g)}\right) \mathrm{d}\boldsymbol{b}_i \,\mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}. \tag{3.5}$$

The probability of allocation to cluster g given the observed data is then by Bayes' theorem proportional to

$$u_{i,g}(\boldsymbol{\theta}) := \mathsf{P}\left[U_i = g \middle| \mathbb{Y}_i, \boldsymbol{\theta}; \mathcal{C}_i\right] \propto w_g \int\!\!\int p\left(\mathbb{Y}_i^{\mathsf{OB}} \,\middle|\, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{\gamma}\right) \cdot \\ \cdot\, p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \,\middle|\, \boldsymbol{b}_i, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}; \mathcal{C}_i\right) \cdot p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\Sigma}^{(g)}\right) \mathrm{d}\boldsymbol{b}_i \,\mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}. \tag{3.6}$$

Direct evaluation for a specific value of $\boldsymbol{\theta}$ requires to approximate the double integral (G times), see Section 7.2.

PBC910 clustering by the *threshold concept* model

We continue in the example in Section 2.4.1 by supposing $G = 2$ latent groups. These groups are assumed to differ in $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ parameters; the predictor remains of the same structure (2.22). The relationships among outcomes will be still described by the matrix $\boldsymbol{\Sigma}$ common to all patients.

Speaking of the matrix $\boldsymbol{\Sigma}$, once we decompose it into standard deviations and correlations the posterior median estimates of the respective elements resemble the previous estimates:

$$\mathsf{diag}\begin{pmatrix} 0.75 \\ 75.50 \\ 1.88 \\ 2.05 \end{pmatrix} \begin{pmatrix} 1.00 & -0.43 & 0.56 & 0.41 \\ -0.43 & 1.00 & -0.36 & -0.23 \\ 0.56 & -0.36 & 1.00 & 0.40 \\ 0.41 & -0.23 & 0.40 & 1.00 \end{pmatrix} \mathsf{diag}\begin{pmatrix} 0.75 \\ 75.50 \\ 1.88 \\ 2.05 \end{pmatrix},$$

although the standard deviations for random effects have decreased. This suggests that some heterogeneity previously captured by the random effects is now captured by the division into the two groups.

Table 3.1: PBC910 dataset, $G = 2$. Posterior medians of the *threshold concept* model parameters including 95% equal-tailed credible intervals.

Parameter	g	Numeric outcomes				Binary outcome		Ordinal outcome	
		Log(bilirubin)		Platelet count		Hepatomegaly		Edema	
β_0	1	1.61	(0.77; 2.45)	297.59	(211.01; 389.29)	−0.89	(−2.98; 1.17)	−4.77	(−7.66; −2.31)
	2	0.33	(−0.32; 0.88)	231.50	(166.86; 292.97)	0.47	(−1.33; 2.41)	−3.73	(−5.99; −1.71)
β_A	1	−0.16	(−0.33; 0.01)	2.35	(−16.36; 20.15)	0.21	(−0.22; 0.63)	0.42	(−0.07; 0.96)
	2	−0.07	(−0.18; 0.07)	3.65	(−8.46; 16.32)	−0.21	(−0.60; 0.16)	0.37	(−0.02; 0.79)
β_M	1	−0.60	(−2.32; 1.17)	32.51	(−181.38; 244.62)	−1.52	(−5.80; 2.66)	−1.14	(−6.64; 4.35)
	2	−0.35	(−1.70; 0.98)	18.01	(−133.28; 168.41)	−0.54	(−5.26; 4.43)	−0.96	(−5.93; 4.13)
$\beta_{A:M}$	1	0.21	(−0.14; 0.55)	−9.21	(−50.41; 33.64)	0.35	(−0.51; 1.24)	−0.15	(−1.26; 0.87)
	2	0.14	(−0.09; 0.38)	−13.37	(−39.98; 13.72)	0.38	(−0.48; 1.22)	0.14	(−0.73; 0.99)
β_1	1	−0.19	(−0.43; 0.06)	−48.90	(−83.42; −15.79)	0.13	(−0.80; 1.00)	0.49	(−0.55; 1.59)
	2	−0.13	(−0.24; −0.02)	−19.45	(−33.95; −5.89)	−0.19	(−0.94; 0.56)	−0.60	(−1.39; 0.20)
β_2	1	0.54	(0.16; 0.91)	8.62	(−42.34; 62.85)	0.52	(−0.68; 1.69)	1.90	(0.38; 3.41)
	2	0.10	(−0.07; 0.28)	−30.34	(−50.62; −9.63)	0.26	(−0.84; 1.30)	0.55	(−0.73; 1.79)
β_3	1	−0.04	(−0.53; 0.46)	−49.28	(−115.96; 20.17)	0.11	(−1.49; 1.78)	1.62	(−0.29; 3.52)
	2	0.13	(−0.08; 0.35)	−37.22	(−63.61; −10.71)	0.19	(−1.27; 1.63)	0.15	(−1.55; 1.78)
$\sigma = \tau^{-\frac{1}{2}}$	1	0.51	(0.46; 0.57)	70.63	(64.49; 78.55)	1		1	
	2	0.26	(0.24; 0.28)	30.40	(26.66; 35.21)	1		1	
γ_1		-		-		-		2.30	(1.95; 2.72)

The difference among the groups is captured by quantile estimates of the posterior distribution of the cluster-specific parameters in Table 3.1. Notably, the first obvious difference lies in the variability of the error terms described by σ parameters. The difference in the effects of the covariates is also captured in Figure 3.2. According to this model, patients in the first cluster (red) reach high values of numeric outcomes and have higher proportion of the hepatomegaly cases, while patients from the second cluster (turquoise) attain much lower values. We again witness that the effect of age changes with the gender of the patient.

To cluster the patients we used the sampled allocation indicators U_i to estimate the posterior probability of allocation by relative frequencies. We have applied clustering rule (P1), where the highest probability had to overcome the threshold of 0.6 to be convincingly clustered to the corresponding cluster. By this rule we have divided the patients to the red group of 94 patients (36.15%, 13 males, 81 females) and the turquoise cluster of 152 patients (58.46%, 14 males, 138 females); the rest 14 (5.38%) females remained unclassified.

For the analysis, only the data from the first 910 days have been used to imitate the situation of knowing only limited amount of data to create a prognosis. However, the original data contain information beyond the 910 days. We have taken the last observation of each patient and evaluated the survival distribution after the day 910. The resulted Kaplan-Meier estimates are depicted in Figure 3.4a, where we can clearly see that the red cluster (of high bilirubin and platelet count) appears to have much worse prognosis with regards to the survival. Hence, we could use this model to sort patients according to the available data into one of these two groups: one safe with better prospects for survival in the future, while the patients sorted into the other cluster should be considered as people with much higher risk of death.

3.3.2 MBC for the GLMM-based model

Similarly to the *threshold concept* model we consider all unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{c}$ and $\boldsymbol{\Sigma}$ of the *GLMM-based* model (Section 2.4.2) to be cluster-specific and we use the same notation $\boldsymbol{\beta}^{(g)}$, $\boldsymbol{\tau}^{(g)}$, $\boldsymbol{c}^{(g)}$ and $\boldsymbol{\Sigma}^{(g)}$ to denote the values specific to cluster $g \in \{1, \dots, G\}$. Yet, we still have the freedom to choose which parameters would stay common to all clusters. The implementation for this model is more advanced and allows even for selecting specific elements of the fixed effects $\boldsymbol{\beta}$ as group-specific (see Section 7.1). Here, only the case of all fixed effects being group-specific is considered. Symbols $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{c}$ and $\boldsymbol{\Sigma}$ from now on unite all the incarnations of the respective parameters, i.e. $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}^{(g)}, g = 1, \dots, G\right\}$, $\boldsymbol{\tau} = \left\{\boldsymbol{\tau}^{(g)}, g = 1, \dots, G\right\}$, $\boldsymbol{c} = \left\{\boldsymbol{c}^{(g)}, g = 1, \dots, G\right\}$, $\boldsymbol{\Sigma} = \left\{\boldsymbol{\Sigma}^{(g)}, g = 1, \dots, G\right\}$, respectively. Altogether, in the notation from the previous section: $\boldsymbol{\zeta} = \emptyset$, $\boldsymbol{\zeta}^{(g)} = \left\{\boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{c}^{(g)}, \boldsymbol{\Sigma}^{(g)}\right\}$ and $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{c}, \boldsymbol{\Sigma}\}$.

Except for the allocation indicators U_i, the only latent variables are the random effects $\mathcal{L}_i = \boldsymbol{b}_i$. The pdf (3.3) for $\mathbb{Y}_i$ under a mixture of the *GLMM-based* models (2.24) can be rewritten to

$$p\left(\mathbb{Y}_i|\boldsymbol{\theta};\mathcal{C}_i\right) = \sum_{g=1}^{G} w_g \int \prod_{r\in\mathcal{R}} \prod_{j=1}^{n_i} p_{\mathsf{t}(r)}\left(Y_{i,j}^r \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r^{(g)}, \tau_r^{(g)}, \boldsymbol{c}_r^{(g)}; \mathcal{C}_{i,j}\right) \cdot p\left(\boldsymbol{b}_i \middle| \boldsymbol{\Sigma}^{(g)}\right) \,\mathrm{d}\boldsymbol{b}_i. \tag{3.7}$$

The probability of allocation to cluster g given the observed data is then by Bayes' theorem proportional to

$$\begin{aligned} u_{i,g}(\boldsymbol{\theta}) := \mathsf{P}\left[U_i = g|\mathbb{Y}_i, \boldsymbol{\theta}; \mathcal{C}_i\right] \propto \\ w_g \int \prod_{r\in\mathcal{R}} \prod_{j=1}^{n_i} p_{\mathsf{t}(r)}\left(Y_{i,j}^r \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r^{(g)}, \tau_r^{(g)}, \boldsymbol{c}_r^{(g)}; \mathcal{C}_{i,j}\right) \cdot p\left(\boldsymbol{b}_i \middle| \boldsymbol{\Sigma}^{(g)}\right) \,\mathrm{d}\boldsymbol{b}_i. \end{aligned} \tag{3.8}$$

Direct evaluation for a specific value of $\boldsymbol{\theta}$ requires to approximate the integral with respect to the random effects $\boldsymbol{b}_i$ (G times), see Section 7.3.

`PBC910` clustering by the *GLMM-based* model

The analysis is analogous to the one with the *threshold concept* model, but a Poisson log-linear mixed model is assumed for the *platelet count*. We suppose $G = 2$ latent groups, each of which follows a model described in Section 2.4.2. The groups differ in the fixed effects $\boldsymbol{\beta}$, precision parameter τ for the *bilirubin* outcome and the ordered intercepts $\boldsymbol{c}$ for *seriousness of edema*. The variance matrix $\boldsymbol{\Sigma}$ describing the associations among random intercepts is kept common to all clusters and was estimated (by posterior median) to be decomposed into

$$\mathsf{diag}\begin{pmatrix}0.88\\0.35\\3.19\\3.18\end{pmatrix} \begin{pmatrix} 1.00 & -0.13 & 0.54 & 0.33 \\ -0.13 & 1.00 & -0.28 & -0.23 \\ 0.54 & -0.28 & 1.00 & 0.35 \\ 0.33 & -0.23 & 0.35 & 1.00 \end{pmatrix} \mathsf{diag}\begin{pmatrix}0.88\\0.35\\3.19\\3.18\end{pmatrix}$$

which resembles the results from Section 2.4.2.

Table 3.2: PBC910 dataset, $G = 2$. Posterior medians of the *GLMM-based* model parameters including 95% equal-tailed credible intervals.

Parameter	g	Numeric outcome		Count outcome		Binary outcome		Ordinal outcome	
		Log(bilirubin)		Platelet count		Hepatomegaly		Edema	
β_0	1	1.42	(0.49;2.42)	5.63	(5.30; 6.13)	2.37	(−1.38; 6.29)	-	
	2	0.65	(−0.13;1.40)	5.49	(5.18; 5.78)	−1.49	(−4.89; 1.53)	-	
β_A	1	−0.17	(−0.37;0.02)	−0.01	(−0.11; 0.04)	−0.46	(−1.26; 0.30)	0.64	(−0.19; 1.63)
	2	−0.10	(−0.25;0.04)	0.01	(−0.05; 0.07)	0.14	(−0.46; 0.82)	0.80	(0.10; 1.61)
β_M	1	−1.27	(−4.57;2.04)	−0.11	(−1.39; 1.18)	−1.45	(−14.83;11.42)	−1.13	(−5.20; 2.57)
	2	0.48	(−2.19;3.60)	0.46	(−0.74; 1.90)	0.10	(−11.66;12.66)	−3.79	(−8.35;−0.42)
$\beta_{A:M}$	1	0.28	(−0.28;0.82)	−0.01	(−0.22; 0.19)	0.73	(−1.38; 2.96)	0.21	(−2.17; 2.70)
	2	0.03	(−0.65;0.60)	−0.11	(−0.42; 0.16)	0.07	(−2.75; 2.71)	0.13	(−4.95; 8.67)
β_1	1	−0.09	(−0.27;0.08)	−0.32	(−0.36;−0.29)	0.43	(−1.18; 1.98)	−0.57	(−2.06; 0.91)
	2	−0.13	(−0.28;0.02)	0.00	(−0.03; 0.02)	−0.49	(−1.70; 0.70)	−0.67	(−2.27; 0.88)
β_2	1	0.00	(−0.27;0.27)	−0.28	(−0.33;−0.22)	−0.90	(−3.41; 1.65)	1.31	(−0.82; 3.36)
	2	0.35	(0.13;0.58)	0.12	(0.08; 0.16)	1.30	(−0.49; 3.07)	2.18	(−0.11; 4.59)
β_3	1	0.44	(0.11;0.77)	−0.65	(−0.73;−0.58)	3.02	(−0.50; 6.64)	1.97	(−0.44; 4.42)
	2	0.07	(−0.21;0.35)	0.07	(0.02; 0.11)	−1.22	(−3.64; 1.00)	−0.55	(−3.94; 2.42)
$\sigma = \tau^{-\frac{1}{2}}$	1	0.38	(0.35;0.42)	-		-		-	
	2	0.37	(0.35;0.40)	-		-		-	
c_0	1	-		-		-		3.02	(2.02; 4.09)
	2	-		-		-		3.42	(2.39; 4.72)
c_1	1	-		-		-		6.62	(5.31; 8.09)
	2	-		-		-		7.56	(6.14; 9.46)

Table 3.2 contains quantile descriptions of the posterior distribution of group-specific model parameters. Many of these parameters appear to be similar in both discovered clusters, which could be adjusted by a detailed model specification. However, the most crucial parameter to distinguish the clusters are the spline coefficients for the *platelet count*. Looking at Figure 3.3, we immediately see that the *platelet count* of patients within the red cluster steeply decreases in time, while it is slowly increasing within the other turquoise cluster. Moreover, men have rather negative effect of age on the *platelet count*. The difference in the effect of age between genders is a lot more striking in the case of *bilirubin*, which agrees with both the exploratory analysis (Figure 1.2) and the *threshold concept* clustering (Figure 3.2).

Similarly as before, we applied the (P1) rule for classification of patients into clusters leaving one man and 6 women unclassified. Comparing the survival curves in Figure 3.4, we see that the new red cluster (41.15%, 14 males and 93 females) and the turquoise cluster (56.15%, 12 males and 134 females) play the very analogous role as in the *threshold concept* model. Yet, the clusters are more balanced and the difference in survival is not that noticeable. This is, perhaps, due to higher focus placed on the evolution of *platelet count* in time.

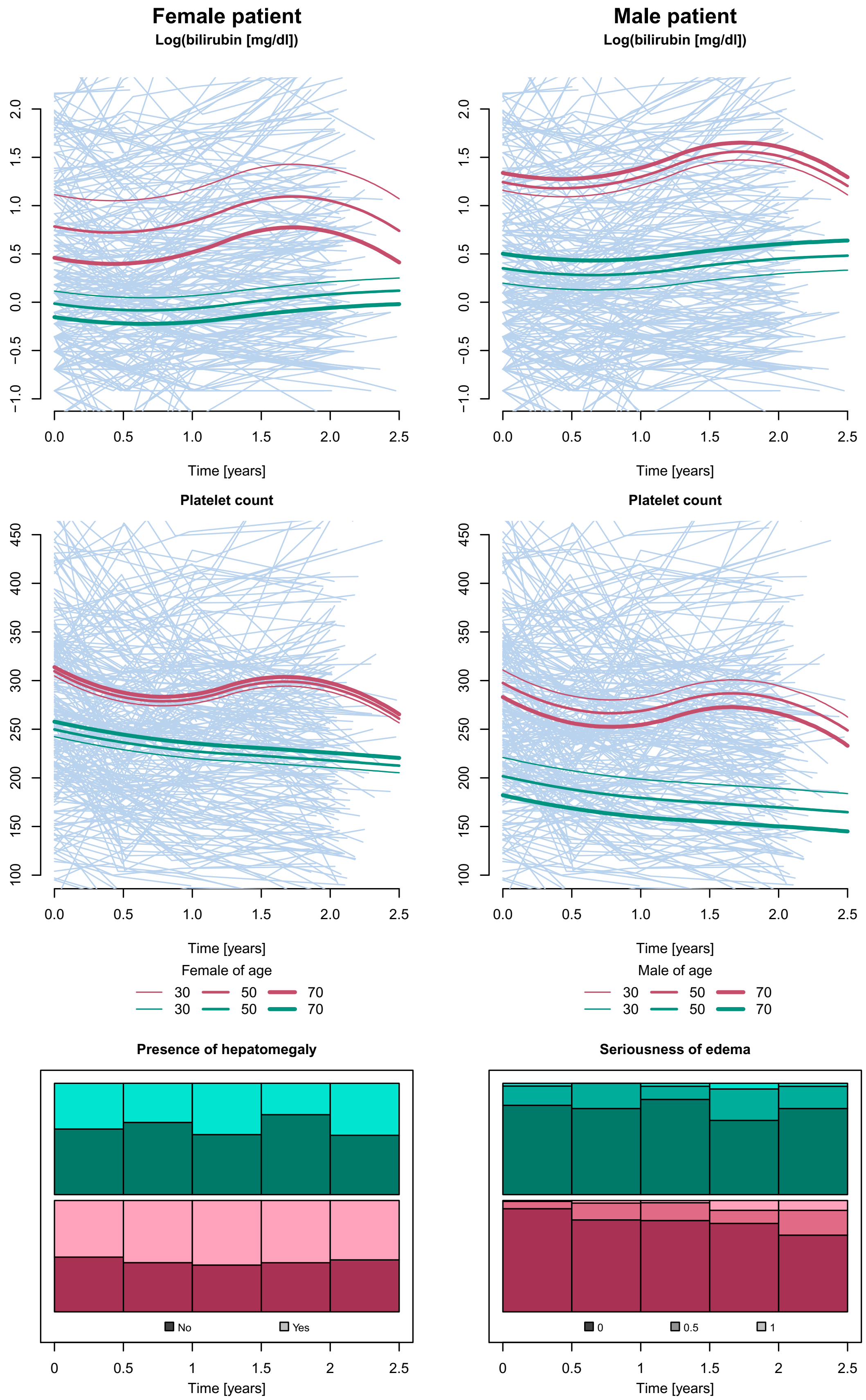


Figure 3.2: PBC910 dataset. *Threshold concept* model, $G = 2$. Estimated group-specific (red ($g = 1$) and turquoise ($g = 2$)) spline curves for patients of different ages for males and females separately by posterior median. Proportions of categorical outcomes with respect to time separately in each cluster.

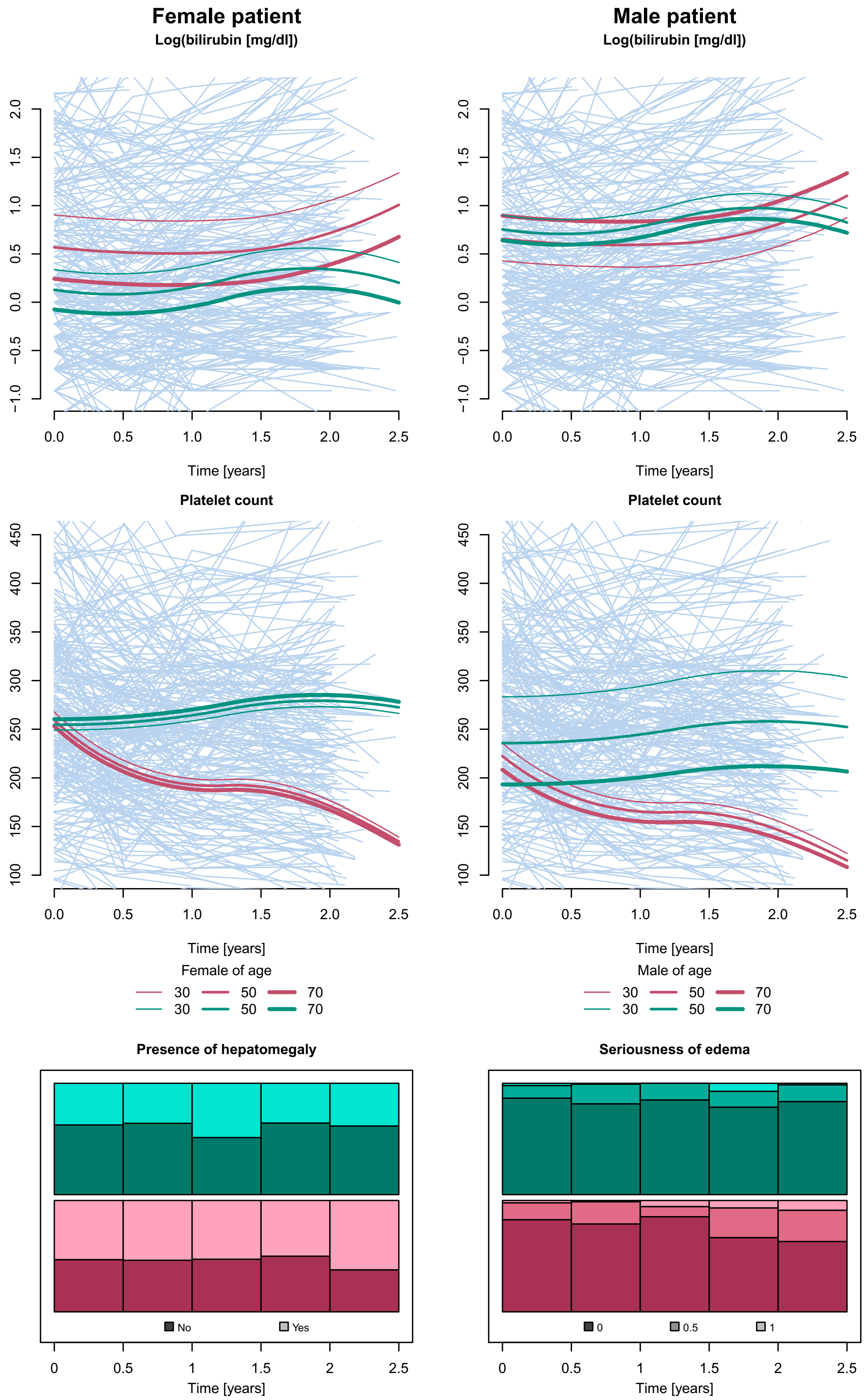
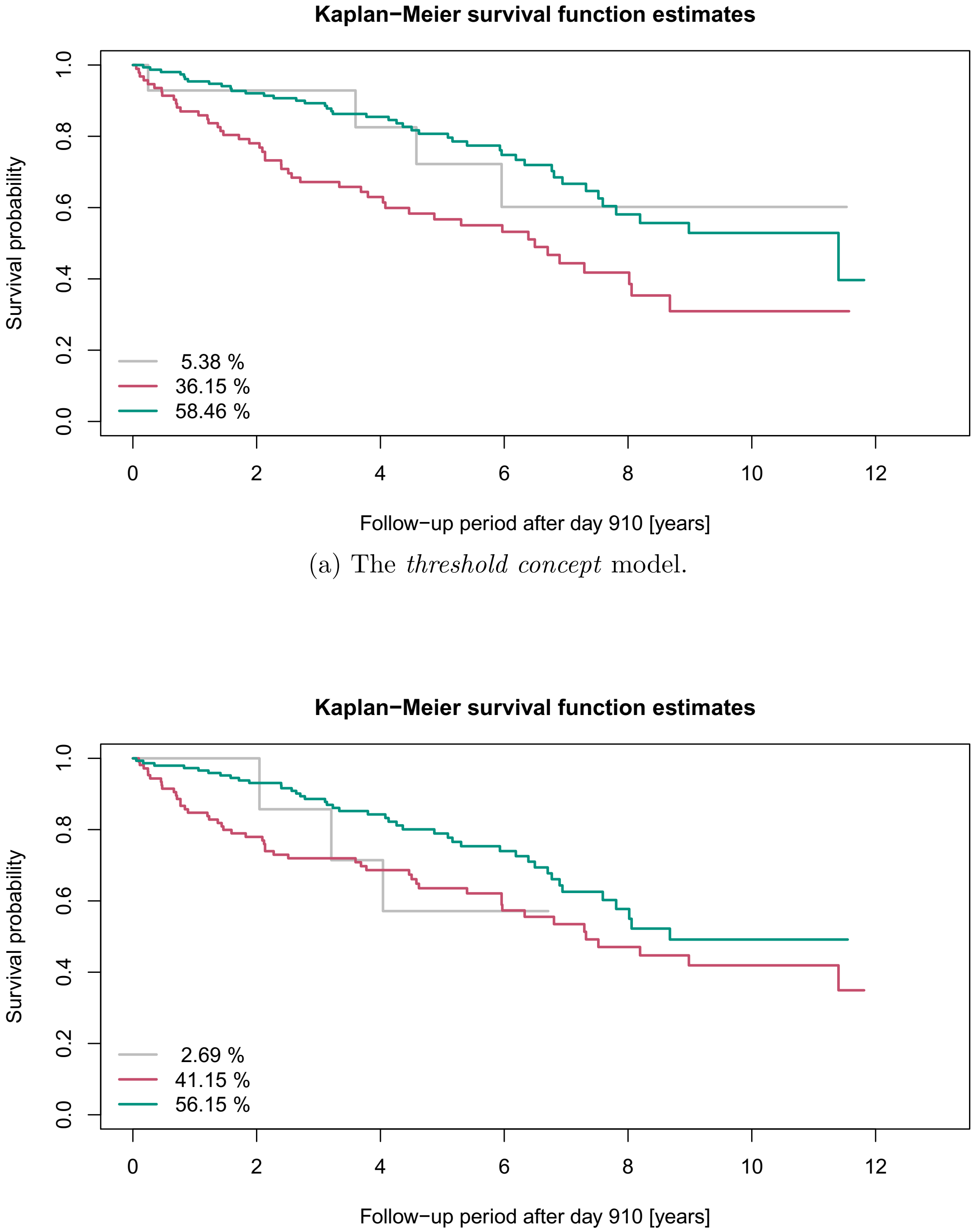


Figure 3.3: `PBC910` dataset. *GLMM-based* model, $G = 2$. Estimated group-specific (red ($g = 1$) and turquoise ($g = 2$)) spline curves for patients of different ages for males and females separately by posterior median. Proportions of categorical outcomes with respect to time separately in each cluster.



(a) The *threshold concept* model.

(b) The *GLMM-based* model.

Figure 3.4: `PBC910` dataset. Kaplan-Meier survival function estimates after the day 910 of $n = 260$ patients clustered into the $G = 2$ discovered groups (red ($g = 1$) and turquoise ($g = 2$)).

4. Bayesian setting and MCMC

In previous chapters we introduced two classes of models for clustering and modelling longitudinal mixed-type data. However, so far we avoided the process of estimating these mixture models. Let us first discuss the possibilities.

Since our data is comprised of n independent blocks (units) following the same model, we can easily form the observed log-likelihood for the set of unknown parameters $\boldsymbol{\theta}$:

$$\ell\left(\boldsymbol{\theta}|\mathbb{Y};\mathcal{C}\right) = \sum_{i=1}^{n} \ell(\mathbb{Y}_i|\boldsymbol{\theta};\mathcal{C}_i), \tag{4.1}$$

where $\ell(\mathbb{Y}_i|\boldsymbol{\theta};\mathcal{C}_i)$ is log-transformed (3.5) or (3.7) for the *threshold concept* model or the *GLMM-based* model, respectively. Nevertheless, since the model is built up on unobserved latent variables (allocation indicators, random effects, latent numeric outcomes), the log-likelihood (4.1) sums up integrals over the latent instances. One could approximate the individual integrals using methods from Chapter 7 and combine them together, which would, however, require sensitivity analysis to guarantee the desired precision. Not to mention that evaluation of the integral is only the first step; an efficient methodology for the subsequent optimisation would have to be proposed, designed and tested.

EM-algorithm (Dempster et al., 1977) is designed to overcome this issue by working with the complete likelihood which does not contain integrals over the latent data since they are treated as observed. Yet, the derivation of the algorithm still requires to perform some integration within the E-step. Bruckers et al. (2016) estimate their clustering procedure for multivariate longitudinal data via EM-algorithm, however, for normally distributed outcomes only. The mixed-type data present a much more difficult challenge, hence, we will rather face the problem of estimation by the Bayesian approach accompanied by MCMC instead.

Bayesian approach allows us to fully exploit hierarchical structures of our models. Regardless of the model complexity, these methods elegantly avoid necessary integrations in a unified way. Moreover, carefully chosen prior distribution of the unknown parameters regularizes the likelihood to elegantly avoid maximisation difficulties caused by unit-specific effects. The clustering itself is then based on the posterior distribution of the individual group probabilities (3.6) or (3.8) and not only on a single point estimate. The Bayesian approach to MBC has been successfully used by Frühwirth-Schnatter (2011) and later by Frühwirth-Schnatter et al. (2012, 2018) to cluster discrete panel data and by Komárek and Komárková (2013) to cluster longitudinal biomedical markers from `PBC910` of a different type.

4.1 Bayesian principles and inference

We digress a bit to provide the reader with the basic principles of the Bayesian approach and to set the necessary terminology.

From now on, we treat any unknown element of our statistical model as random. It is assumed to follow a certain *prior* distribution, for example, the pdf for unknown parameters $\boldsymbol{\theta}$ is denoted by $p(\boldsymbol{\theta})$. It expresses analyst's prior beliefs about the value of the element before the data are gathered and known. Once the data are observed we may update our prior belief into the *posterior*. Strictly speaking, the posterior distribution is conditional distribution of $\boldsymbol{\theta}$ given the observed data. By Bayes' theorem we know that the posterior pdf is proportional to the product of the model contribution with the prior belief:

$$p(\boldsymbol{\theta} \,|\, \mathbb{Y}; \mathcal{C}) \propto \underbrace{p(\mathbb{Y} \,|\, \boldsymbol{\theta}; \mathcal{C})}_{\text{likelihood}} \cdot \underbrace{p(\boldsymbol{\theta})}_{\text{prior}}, \tag{4.2}$$

where the model is specified only for data $\mathbb{Y}$ given covariates $\mathcal{C}$, for which no model is assumed.

The prior distribution could be based on some historical expert knowledge or left completely vague. In the latter case, all values are equally plausible ($p(\boldsymbol{\theta}) \propto 1$) and the model specification in the form of likelihood is the only source of information about the unknown parameter. This is why such a prior is called *non-informative*.

On the other hand, an *informative* prior prefers some values over the other. The key issue here is to express how precisely are we certain about our prior belief. This can be expressed by appropriate setting of the hyperparameters of the prior distribution. This is also where the critique of Bayesian principles stems from since one can (even unintentionally) overshadow the information provided by the gathered data and shape the result to any desired value. To avoid such abuse of Bayesian methods we prefer priors very close to the non-informative prior (*flat* prior of low precision) but still informative enough to regularize potential issues of the model contribution.

The inference about unknown parameters then comes from the posterior distribution which can be described by any distributional characteristic. For example, characteristics of location, such as posterior mean, median or mode, are suitable replacements for the respective point estimators in the classical frequentist statistics. The equivalent of 95% confidence region for a parameter is the *credible* region – a set of values that covers exactly 95% of the posterior mass. There are several different construction methods for *credible regions*. The *highest posterior density* (HPD) credible region gathers all values for which the posterior density is higher than a threshold such that the desired coverage is still achieved. In the case of one dimensional parameter, 0.025 and 0.975 posterior quantiles declare the bounds of the so called *equal-tailed* (ET) credible interval. One of the advantages of the credible intervals in general is that they take into account any potential skewness or multimodality of the posterior distribution, while the confidence intervals of the form "*point estimate* ± *standard error*" ignore such possibility.

4.2 Prior distribution settings

We consider rather standard prior distributions of primary model parameters $\boldsymbol{\theta}$ used in the context of hierarchical models. In particular, we assume that the prior distribution is decomposed into

$$p(\boldsymbol{\theta}, \mathbb{Q}) = \underbrace{p(\boldsymbol{w})}_{(4.11)} \underbrace{p(\boldsymbol{\gamma})}_{(4.13)} \underbrace{p(\boldsymbol{\beta} \,|\, \boldsymbol{\tau})\, p(\boldsymbol{\tau})}_{(4.5)} \underbrace{p(\boldsymbol{\Sigma}|\mathbb{Q})\, p(\mathbb{Q})}_{(4.9),(4.10)} \tag{4.3}$$

for the *threshold concept* model and into

$$p(\boldsymbol{\theta},\, e_0,\, \mathbb{Q}) = \underbrace{p(\boldsymbol{w}|e_0)}_{(4.11)}\, \underbrace{p(e_0)}_{(4.12)}\quad \underbrace{p(\boldsymbol{c})}_{(4.14),(4.15)}\quad \underbrace{p(\boldsymbol{\beta}\,\big|\,\boldsymbol{\tau})\, p(\boldsymbol{\tau})}_{(4.7)}\, \underbrace{p(\boldsymbol{\Sigma}|\mathbb{Q})\, p(\mathbb{Q})}_{(4.9),(4.10)} \tag{4.4}$$

for the *GLMM-based* model, where $\mathbb{Q}$ and e_0 are additional hyperparameters considered to be random. Both models share similar ideas for the elements of factorization in (4.3) and (4.4) with some minor differences.

In the following, we discuss the prior specification only under the group-specificity of all possible parameters. In case some parameters are required to be common to all groups, the priors are specified in analogous way after appropriate modifications. Though, the hierarchy has to be respected; a parameter shared by all clusters cannot be generated by a distribution given by a group-specific parameter.

Fixed effects $\boldsymbol{\beta}$ and precisions $\boldsymbol{\tau}$

The regression coefficients for numeric outcomes $\boldsymbol{\beta}_r^{(g)} = \left(\beta_{r,1}^{(g)},\, \ldots,\, \beta_{r,d_r^{\mathsf{F}}}^{(g)}\right)$, $r \in \mathcal{R}^{\mathsf{Num}}$, $g = 1, \ldots, G$, are assumed to be a-priori independent and follow a conjugate normal distribution in combination with the precision parameter $\tau_r^{(g)}$, that is $\mathsf{N}\left(\beta_{0,r,j},\, \left(\tau_r^{(g)}\right)^{-1} d_{j,j}^r\right)$ where $\beta_{0,r,j}$ and $d_{j,j}^r$ are fixed hyperparameters. In the applications, these hyperparameters are frequently set equal to 0 and 10, respectively, to induce a flat prior. In a vector notation, $\boldsymbol{\beta}_r^{(g)} \sim \mathsf{N}_{d_r^{\mathsf{F}}}\left(\boldsymbol{\beta}_{0,r},\, \left(\tau_r^{(g)}\right)^{-1}\mathbb{D}_r\right)$, where $\boldsymbol{\beta}_{0,r} = \left(\beta_{0,r,1},\, \ldots,\, \beta_{0,r,d_r^{\mathsf{F}}}\right)$ and $\mathbb{D}_r = \mathsf{diag}\left(d_{j,j}^r,\, j = 1, \ldots, d_r^{\mathsf{F}}\right)$.

The precision parameters $\tau_r^{(g)}$ for numeric outcomes are assumed to follow independent Gamma priors $\tau_r^{(g)} \sim \Gamma(a_\tau,\, b_\tau)$ with shape $a_\tau > 0$ and rate $b_\tau > 0$. For calculations in the later applications, we often use $a_\tau = b_\tau = 1$.

The regression coefficients for the count, binary and ordinal, i.e. $\beta_{r,j}^{(g)}, r \in \mathcal{R}^{\mathsf{Poi}} \cup \mathcal{R}^{\mathsf{Bin}} \cup \mathcal{R}^{\mathsf{Ord}}$ are also assumed to be a-priori independent and follow an analogous normal distribution $\mathsf{N}\left(\beta_{0,r,j},\, d_{j,j}^r\right)$, where, however, no precision parameter $\boldsymbol{\tau}$ is involved. Equivalently, in a similar vector notation, $\boldsymbol{\beta}_r^{(g)} \sim \mathsf{N}_{d_r^{\mathsf{F}}}\left(\boldsymbol{\beta}_{0,r},\, \mathbb{D}_r\right)$. Priors are analogously set for effects of the general categorical outcomes, where we have to moreover distinguish effects $\boldsymbol{\beta}_{r,k}^{(g)}$ for different levels $k \in \{1, \ldots, K^r - 1\}$, i.e. $\boldsymbol{\beta}_{r,k}^{(g)} \sim \mathsf{N}_{d_r^{\mathsf{F}}}\left(\boldsymbol{\beta}_{0,r},\, \mathbb{D}_r\right)$.

Altogether, we have

$$p(\boldsymbol{\beta}|\boldsymbol{\tau})p(\boldsymbol{\tau}) = \prod_{g=1}^{G}\prod_{r\in\mathcal{R}^{\mathsf{Num}}}\left[\varphi\left(\boldsymbol{\beta}_r^{(g)};\, \boldsymbol{\beta}_{0,r},\, \left(\tau_r^{(g)}\right)^{-1}\mathbb{D}_r\right) p\left(\tau_r^{(g)}\,\Big|\, a_\tau, b_\tau\right)\right]\cdot \\ \cdot \prod_{g=1}^{G}\prod_{r\in\mathcal{R}^{\mathsf{OB}}}\varphi\left(\boldsymbol{\beta}_r^{(g)};\, \boldsymbol{\beta}_{0,r},\, \mathbb{D}_r\right) \tag{4.5}$$

and

$$\begin{aligned}\ell(\boldsymbol{\beta}|\boldsymbol{\tau}) + \ell(\boldsymbol{\tau}) = \text{const.} + \sum_{g=1}^{G}\sum_{r\in\mathcal{R}^{\mathsf{Num}}} \left(a_\tau + \frac{d_r^{\mathsf{F}}}{2} - 1\right)\log\left(\tau_r^{(g)}\right) - \\ - \sum_{g=1}^{G}\sum_{r\in\mathcal{R}^{\mathsf{Num}}} \tau_r^{(g)}\left[b_\tau + \frac{1}{2}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)\right] - \\ - \frac{1}{2}\sum_{g=1}^{G}\sum_{r\in\mathcal{R}^{\mathsf{OB}}} \left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right) \quad (4.6)\end{aligned}$$

for the *threshold concept* model. Analogously, for the *GLMM-based* model:

$$\begin{aligned}p(\boldsymbol{\beta}|\boldsymbol{\tau})p(\boldsymbol{\tau}) = \prod_{g=1}^{G}\prod_{r\in\mathcal{R}^{\mathsf{Num}}}\left[\varphi\left(\boldsymbol{\beta}_r^{(g)};\, \boldsymbol{\beta}_{0,r},\, \left(\tau_r^{(g)}\right)^{-1}\mathbb{D}_r\right) p\left(\tau_r^{(g)}\,\middle|\, a_\tau, b_\tau\right)\right]\cdot \\ \cdot\prod_{g=1}^{G}\prod_{r\in\mathcal{R}^{\mathsf{Poi}}\cup\mathcal{R}^{\mathsf{Bin}}\cup\mathcal{R}^{\mathsf{Ord}}} \varphi\left(\boldsymbol{\beta}_r^{(g)};\, \boldsymbol{\beta}_{0,r},\, \mathbb{D}_r\right)\prod_{g=1}^{G}\prod_{r\in\mathcal{R}^{\mathsf{Cat}}}\prod_{k=1}^{K^r-1}\varphi\left(\boldsymbol{\beta}_{r,k}^{(g)};\, \boldsymbol{\beta}_{0,r},\, \mathbb{D}_r\right) \quad (4.7)\end{aligned}$$

and

$$\begin{aligned}\ell(\boldsymbol{\beta}|\boldsymbol{\tau}) + \ell(\boldsymbol{\tau}) = \text{const.} + \sum_{g=1}^{G}\sum_{r\in\mathcal{R}^{\mathsf{Num}}} \left(a_\tau + \frac{d_r^{\mathsf{F}}}{2} - 1\right)\log\left(\tau_r^{(g)}\right) - \\ - \sum_{g=1}^{G}\sum_{r\in\mathcal{R}^{\mathsf{Num}}} \tau_r^{(g)}\left[b_\tau + \frac{1}{2}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)\right] - \\ - \frac{1}{2}\sum_{g=1}^{G}\sum_{r\in\mathcal{R}^{\mathsf{Poi}}\cup\mathcal{R}^{\mathsf{Bin}}\cup\mathcal{R}^{\mathsf{Ord}}} \left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right) - \\ - \frac{1}{2}\sum_{g=1}^{G}\sum_{r\in\mathcal{R}^{\mathsf{Cat}}}\sum_{k=1}^{K^r-1} \left(\boldsymbol{\beta}_{r,k}^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_{r,k}^{(g)} - \boldsymbol{\beta}_{0,r}\right). \quad (4.8)\end{aligned}$$

Covariance matrix $\boldsymbol{\Sigma}$ for random effects

The covariance matrices $\boldsymbol{\Sigma}^{(g)}$ of the random effects $\boldsymbol{b}_i$ are general positive-definite matrices. We impose a Wishart prior on the inverse covariance matrices $\boldsymbol{\Sigma}^{-(g)} := \left(\boldsymbol{\Sigma}^{(g)}\right)^{-1}$ to preserve conjugacy. The parameters of the Wishart prior are the scale matrix $\mathbb{Q}$ and the number of degrees of freedom $\nu_0 \geq d^{\mathsf{R}}$. To avoid selecting a specific value for the scale matrix and aiming at obtaining a weakly informative prior for the covariance matrices, we also assume a prior for the scale matrix $\mathbb{Q}$ while keeping the number of degrees of freedom $\nu_0 \geq d^{\mathsf{R}}$ fixed. Again a Wishart prior is assumed for the inverse scale matrix $\mathbb{Q}^{-1}$. For this prior, fixed values are selected for the scale matrix and the number of degrees of freedom ν_1. In our applications, we use $\nu_0 = \nu_1 = d^{\mathsf{R}} + 1$ and a diagonal matrix for the scale matrix given by $\mathbb{D}_{\mathbb{Q}} = 100 \cdot \mathbb{I}_{d^{\mathsf{R}}}$.

The corresponding pdfs in terms of the inverse matrices can be expressed as

$$p\left(\boldsymbol{\Sigma}^{-(g)}\,\middle|\,\mathbb{Q};\nu_0\right) \propto \left|\boldsymbol{\Sigma}^{-(g)}\right|^{\frac{\nu_0 - d^{\mathsf{R}} - 1}{2}}\left|\mathbb{Q}^{-1}\right|^{\frac{\nu_0}{2}}\exp\left\{-\frac{1}{2}\,\mathsf{Tr}\left[\mathbb{Q}^{-1}\boldsymbol{\Sigma}^{-(g)}\right]\right\} \quad (4.9)$$

and

$$p\left(\mathbb{Q}^{-1}\,\middle|\,\mathbb{D}_{\mathbb{Q}};\nu_1\right) \propto \left|\mathbb{Q}^{-1}\right|^{\frac{\nu_1 - d^{\mathsf{R}} - 1}{2}}\exp\left\{-\frac{1}{2}\,\mathsf{Tr}\left[\mathbb{D}_{\mathbb{Q}}^{-1}\mathbb{Q}^{-1}\right]\right\}. \quad (4.10)$$

Cluster allocation probabilities $\boldsymbol{w}$

The vector of marginal allocation probabilities $\boldsymbol{w}$ lives on a simplex, where $0 < w_g < 1$ and $w_1 + \cdots + w_G = 1$. Hence, Dirichlet distribution $\mathsf{Dir}_G(e_1, \ldots, e_G)$ is a very popular choice, where $e_g > 0$ represents *the number of units within the cluster g a priori.* Traditionally, all clusters are given the same weight $0 < e_0 = e_g$ for all $g = 1, \ldots, G$. This symmetric version of Dirichlet distribution is briefly denoted by $\mathsf{Dir}_G(e_0)$. Hence, assuming $\boldsymbol{w} \sim \mathsf{Dir}_G(e_0)$ yields

$$p(\boldsymbol{w}) = \frac{\Gamma\left(\sum\limits_{g=1}^{G} e_g\right)}{\prod\limits_{g=1}^{G} \Gamma(e_g)} \prod_{g=1}^{G} w_g^{e_g - 1} \quad \overset{e_g = e_0}{\propto} \quad \prod_{g=1}^{G} w_g^{e_0 - 1}. \tag{4.11}$$

Frühwirth-Schnatter and Malsiner-Walli (2019) point out that the choice of e_0 is essential for controlling the sparsity of mixture components. High value of e_0 leads to balanced weights $\boldsymbol{w}$ with high probability. However, as Figure 4.1 illustrates in case of $G = 3$ by decreasing e_0 the mass begins to concentrate on the edges and vertices of the equilateral triangle representing the simplex. In general, when e_0 is close to 0, the Dirichlet prior puts a lot of mass on the boundary regions of the simplex and many of the G weights are small a-priori. Then, a sample of allocation indicators $U_1, \ldots, U_n$ generated from multinomial distribution with such extremely unbalanced $\boldsymbol{w}$ may not even contain some of the G possible values. Hence, in such a case the number of *non-empty components*

$$G_+ = G - \sum_{g=1}^{G} \mathbb{1}_{\left(n^{(g)} = 0\right)},$$

where $n^{(g)} = \sum_{i=1}^{n} \mathbb{1}_{(U_i = g)}$, satisfies $G_+ \ll G$ with high probability. This is the fundamental idea behind the *sparse finite mixture* approach (Malsiner-Walli et al., 2016) for estimating the appropriate number of mixture components.

Let us denote by $G_{\max}$ the maximal number of components considered. In case the number of groups G are a-priori known, one would set $G_{\max} = G$ and

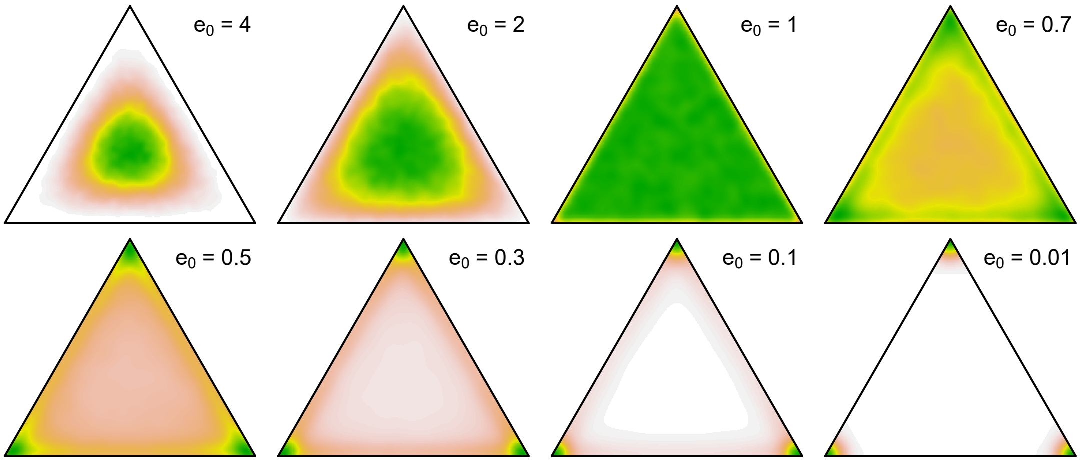


Figure 4.1: Density of a symmetric Dirichlet distribution $\mathsf{Dir}_3(e_0)$ displayed in an equilateral triangle under several choices of parameter e_0.

use a rather large value for e_0. If G is not known in advance, then a sparse finite mixture is supposed by setting a small value for e_0 to estimate G by $G_+ < G_{\max}$. More about the estimation process can be found later in Section 6.3.

To attenuate the influence of a specific choice of e_0, we assign a Gamma prior $e_0|a_e, b_e \sim \Gamma(a_e, b_e)$ with pdf

$$p(e_0|a_e, b_e) = \frac{b_e^{a_e}}{\Gamma(a_e)} e_0^{a_e-1} \exp\{-b_e e_0\} \tag{4.12}$$

and prior expected value $\mathsf{E}(e_0) = a_e/b_e$. As recommended by Frühwirth-Schnatter and Malsiner-Walli (2019), we select the parameters a_e and b_e of the Gamma prior to have a small mean when aiming at sparsity, i.e. $\mathsf{E}(e_0) = a_e/b_e = 0.01$ with $a_e = 1$. In case the number of components G are assumed known and one aims at $G_+ \approx G$, we select the parameters to induce a mean of $\mathsf{E}(e_0) = a_e/b_e = 4$ or directly fix $e_0 = 4$ to avoid sparsity.

Ordered parameters $\boldsymbol{\gamma}$ and $\boldsymbol{c}$

Ordinal outcomes $r \in \mathcal{R}^{\mathsf{Ord}}$ require additional parameter for parametrizing the probabilities. The *threshold concept* model works with ordered thresholds $-\infty = \gamma^r_{-1} < \gamma^r_0 < \cdots < \gamma^r_{K^r-1} = \infty$, while the *GLMM-based* model uses ordered intercepts $-\infty = c_{r,-1} < c_{r,0} < c_{r,1} < \cdots < c_{r,K^r-1} = \infty$.

Considering the thresholds $\boldsymbol{\gamma}^r = (\gamma^r_1, \ldots, \gamma_{K^r-2})^\top$, $r \in \mathcal{R}^{\mathsf{Ord}}$, we first address the identifiability issue. Corresponding parametric space Ω^r is limited to a set of all vectors of ordered values with fixed first threshold $\gamma^r_0 = 0$:

$$\Omega^r = \{\boldsymbol{\gamma} \in \mathbb{R}^{K^r-2} : 0 = \gamma^r_0 < \gamma_1 < \cdots < \gamma_{K^r-2}\}.$$

An improper uniform distribution on Ω^r is assumed for each set of thresholds $\boldsymbol{\gamma}^r$, $r \in \mathcal{R}^{\mathsf{Ord}}$. That is,

$$p(\boldsymbol{\gamma}) = \prod_{r \in \mathcal{R}^{\mathsf{Ord}}} p(\boldsymbol{\gamma}^r) \propto \prod_{r \in \mathcal{R}^{\mathsf{Ord}}} \mathbb{1}_{\Omega^r}(\boldsymbol{\gamma}^r). \tag{4.13}$$

Regarding the ordered intercepts $\boldsymbol{c}_r^{(g)}$ the prior is not specified directly but for transformed quantities. The (K^r-1)-dimensional ordered intercepts $\boldsymbol{c}_r^{(g)} = \left(c^{(g)}_{r,0}, \ldots, c^{(g)}_{r,K^r-2}\right)^\top$ are transformed into $\boldsymbol{\pi}_r^{(g)} = \left(\pi^{(g)}_{r,0}, \ldots, \pi^{(g)}_{r,K^r-1}\right)^\top$:

$$\begin{aligned} \pi^{(g)}_{r,k} &= \mathsf{P}\left[Y^r_{i,j} = k \,\middle|\, \boldsymbol{b}_i = \boldsymbol{0},\, U_i = g,\, \boldsymbol{x}^r_{i,j} = \boldsymbol{0}\right] \\ &= \mathsf{logit}^{-1}\left(c^{(g)}_{r,k}\right) - \mathsf{logit}^{-1}\left(c^{(g)}_{r,k-1}\right), \\ c^{(g)}_{r,k} &= \log\left(\frac{\pi^{(g)}_{r,0} + \cdots + \pi^{(g)}_{r,k}}{\pi^{(g)}_{r,k+1} + \cdots \pi^{(g)}_{r,K^r-1}}\right). \end{aligned} \tag{4.14}$$

The prior distribution is then specified for the probabilities $\boldsymbol{\pi}_r^{(g)}$ adding up to 1 for all outcomes $r \in \mathcal{R}^{\mathsf{Ord}}$ using a product of Dirichlet distributions:

$$p(\boldsymbol{\pi}) \propto \prod_{g=1}^{G} \prod_{r \in \mathcal{R}^{\mathsf{Ord}}} \prod_{k=0}^{K^r-1} \left(\pi^{(g)}_{r,k}\right)^{\alpha_{r,k}-1}, \tag{4.15}$$

where the hyperparameters $\alpha_{r,k}$ are fixed. A value of 1 inducing a uniform distribution on the simplex is used in the later applications.

4.3 Posterior distribution exploration

First, we have to examine Bayes' rule (4.2) more carefully. Assumed models are built upon latent variables $\mathcal{L}$ such as random effects $\boldsymbol{b}_i$, group allocations U_i, etc. Therefore, a more precise version of (4.2) should be

$$p(\boldsymbol{\theta} \,|\, \mathbb{Y}; \mathcal{C}) \propto \; p(\boldsymbol{\theta}) \; \cdot \int p(\mathbb{Y} \,|\, \boldsymbol{\theta}, \, \mathcal{L}; \, \mathcal{C}) \, p(\mathcal{L} \,|\, \boldsymbol{\theta}; \, \mathcal{C}) \, \mathrm{d}\mathcal{L},$$

where the latent data are integrated out to obtain the pure observed likelihood. As mentioned several times throughout the thesis, this is hardly tractable. Hence, we exploit the ideas of *Bayesian data augmentation* (BDA) by Tanner and Wong (1987) while considering all latent quantities as additional model parameters included in the posterior distribution. This includes any potential randomized hyperparameters (collectively denoted by $\mathcal{H}$) as well, e.g. $\mathbb{Q}$ and e_0.

Then, Bayes' rule (4.2) under presence of latent elements $\mathcal{L}$ and randomized hyperparameters $\mathcal{H}$ takes the folowing form:

$$p(\boldsymbol{\theta}, \, \mathcal{L}, \, \mathcal{H} \,|\, \mathbb{Y}; \, \mathcal{C}) \propto \underbrace{p(\mathbb{Y} \,|\, \boldsymbol{\theta}, \, \mathcal{L}; \, \mathcal{C}) \, \cdot \, p(\mathcal{L} \,|\, \boldsymbol{\theta}; \, \mathcal{C})}_{\text{complete likelihood}} \, \cdot \, \underbrace{p(\boldsymbol{\theta} \,|\, \mathcal{H}) \, \cdot \, p(\mathcal{H})}_{\text{prior}}. \qquad (4.16)$$

Due to complexity of our suggested models, we cannot match the posterior of $\boldsymbol{\Psi} = \{\boldsymbol{\theta}, \mathcal{L}, \mathcal{H}\}$ with any commonly used distributional family. Moreover, evaluating the marginals of each random element $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ would be painful. Luckily, Markov chain Monte Carlo (MCMC) methods elegantly evade any problems with integration and provide a direct access to the marginals.

MCMC

MCMC methods will be mentioned here only briefly, more details are provided by Brooks et al. (2011), Robert and Casella (2004) or Hastings (1970).

Given a random sample $\{\boldsymbol{\Psi}^m = \{\boldsymbol{\theta}^m, \mathcal{L}^m, \mathcal{H}^m\}, m = 1, \ldots, M\}$ of size M from the target distribution, we could estimate any of its characteristics with a reasonable precision. This is the way of Monte Carlo. For example, any univariate parametric function $t(\boldsymbol{\Psi})$ satisfying $\mathsf{E}\left[|t(\boldsymbol{\Psi})| \middle| \mathbb{Y}; \mathcal{C}\right] < \infty$ can be estimated by averaging the values $t\left(\boldsymbol{\Psi}^m\right), m = 1, \ldots, M$, by the strong law of large numbers (SLLN). However, how to obtain a random sample from the posterior distribution?

This is where the first 'MC' from MCMC (Markov chain) shines. We will generate a homogeneous Markov chain with the states $\{\boldsymbol{\psi}^m\}$ and the following properties:

1. its stationary distribution is the posterior distribution $\boldsymbol{\Psi}|\mathbb{Y};\mathcal{C}$ and

2. there exists a limiting distribution.

From certain $B > 0$ large enough, the generated states $\{\boldsymbol{\Psi}^{B+m}, m = 1, \ldots, M\}$ can be considered as representatives of the limiting distribution. Knowing that the limiting distribution of a homogeneous Markov chain coincides with the stationary one, these states could be considered 'a sample' from the posterior distribution.

The problem of (often highly) autocorrelated states violates the assumption of independence in the SLLN. Even though the autocorrelations can be substantially reduced by thinning (keeping only every, say, tenth state), usually some form of dependence remains. Fortunately, the ergodic theorem (Robert and Casella, 2004, Theorem 6.63) still enables us to estimate the posterior mean of $t(\boldsymbol{\Psi})$ by the average of values $t\left(\boldsymbol{\Psi}^m\right), m = 1, \ldots, M$ provided the Markov chains is well behaved. The necessary properties are not difficult to secure (Brooks et al., 2011, Section 2.5.4).

One algorithm generating a Markov chain of such properties is the *Gibbs sampler* by Geman and Geman (1984). The states are divided into blocks $\boldsymbol{\psi}$, in our case $\boldsymbol{\psi} \in \{\boldsymbol{w}, \boldsymbol{\beta}_r^{(g)}, \tau_r^{(g)}, \boldsymbol{\gamma}^r, \boldsymbol{c}_r^{(g)}, \boldsymbol{\Sigma}^{(g)}, U_i, \boldsymbol{b}_i, \mathbb{Y}_i^{\star,\mathsf{OB}}, \mathbb{Q}, e_0\}$ of appropriate indices r, i and g. Then each block $\boldsymbol{\psi}$ is sampled from the *full-conditional* distribution $\boldsymbol{\psi}\,|\boldsymbol{\Psi}_{-\psi}, \mathbb{Y}; \mathcal{C}$, where we condition on the observed data and the *last known values* of the rest of the parameters $\boldsymbol{\Psi}_{-\psi} = \boldsymbol{\Psi} \setminus \{\boldsymbol{\psi}\}$. These distributions are easy to work with since we condition even by the latent quantities instead of integrating them. The full-conditional distribution will be briefly denoted by $\boldsymbol{\psi}| \cdots$. Every block $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, one by one, is resampled until a whole new $\boldsymbol{\Psi}^m$ is obtained and is later used to generate the next $\boldsymbol{\Psi}^{m+1}$.

However, it is required that each full-conditional distribution $\boldsymbol{\psi}| \cdots$ belongs to a well known distributional family or at least an efficient sampler for this distribution exists. We have managed to fulfil this requirement with the *threshold concept* model. However, within the *GLMM-based* model the full-conditional distribution of several parameters is hardly tractable. Hence, we replace the problematic steps with a *Metropolis* proposal step. The idea is to propose a new value of $\boldsymbol{\psi}$ from the proposal distribution (multivariate normal) and then with according probability accept the proposal value as a new state, otherwise deny it and remain in the current state. Details on how the appropriate proposal distributions are designed and the calculation of the acceptance probabilities are covered in detail later in Section 7.4.

Label switching problem

Redner and Walker (1984) point out the *label switching problem* which often arises with mixture models. Especially, when the cluster labels are not a priori assigned by a given rule. It comes from the fact that the likelihood as well as the prior and thus the posterior are invariant towards the permutation of cluster labels. During the sampling procedure it may happen that the meaning of the labels $g = 1, \ldots, G$ is switched. Then, cluster-specific parameters may appear to have multi-modal posterior because they consist of modes corresponding to the switched labels.

One can detect such issues by visualization of the traceplots of every parameter, where sudden shifts are symptoms of label switching. With hundreds of parameters this task may be tiresome for human. To automatize the detection, Stephens (2000) proposed a post-sampling procedure which considers all $G!$ permutations of labels for each iteration and ensures that the latent clusters $1, \ldots, G$ have a fixed meaning during the whole sampling procedure. Only after label switching has been addressed one should proceed with inference sensitive to the change of cluster labels such as the estimation of classification probabilities. When sparse finite mixtures are induced, we use more elaborate procedure

(Algorithm 3) designed for such a case by Frühwirth-Schnatter and Malsiner-Walli (2019), which is based on k-means clustering (Hartigan and Wong, 1979) (Algorithm 7). See Section 6.3 for more details.

From applications, it seems that the complexity of our models prevents such problems completely since we have not encountered any. However, to properly judge the convergence to the stationary distribution we sample several chains each started from a different randomly assigned initial values. All chains usually come to the same results, even though some cluster labels have different meanings across chains, see Figure 4.2 for an example. Clearly, chains 2 and 4 share the same cluster interpretation, while chains 1 and 3 have different meaning of $g = 1, 2, 3$. Nevertheless, it is obvious that there exist a permutation yielding the same results. If one aims to use all sampled chains for inference, then the labels have to be unified by appropriate permutation of all cluster-specific parameters in each chain.

Classification probabilities

Probabilities $u_{i,g}(\boldsymbol{\theta})$ defined in (3.6) and (3.8) are viewed as parametric functions of $\boldsymbol{\theta}$. Therefore, their posterior distribution is explored by evaluation of $u_{i,g}\left(\boldsymbol{\theta}^m\right)$, $m \in 1, \ldots, M$, which is computationally expensive, see Sections 7.2 and 7.3.

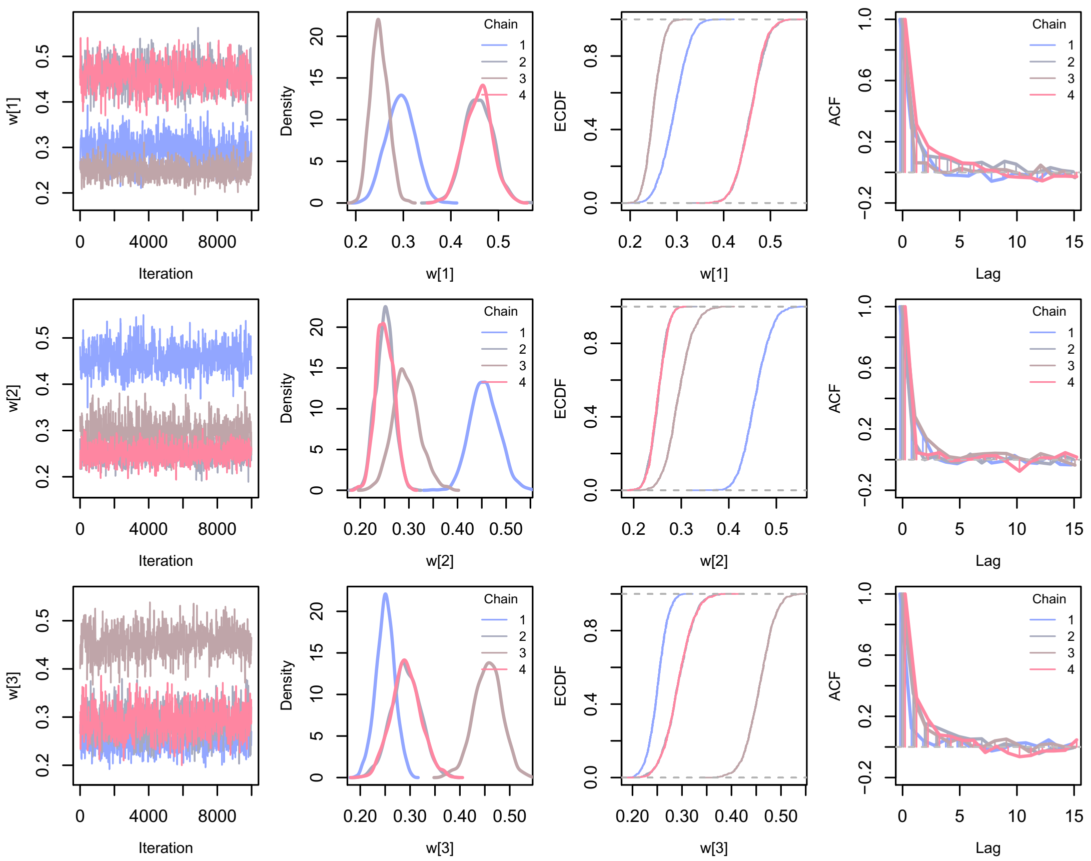


Figure 4.2: Simulated dataset of $G = 3$ clusters, $n = 250$, $n_i = 4$ and true value $\boldsymbol{w} = (0.25, 0.33, 0.42)^\top$ estimated by the *GLMM-based* model. Traceplots, kernel density estimates, empirical cumulative distribution functions (ECDF) and autocorrelation functions (ACF) of four different chains of parameter $\boldsymbol{w}$.

However, when we realize that

$$\mathsf{E}\left[u_{i,g}(\boldsymbol{\theta})\right|\mathbb{Y};\mathcal{C}] = \mathsf{E}\left[\mathsf{P}\left[U_i = g|\boldsymbol{\theta},\mathbb{Y}_i;\mathcal{C}_i\right]\middle|\mathbb{Y};\mathcal{C}\right] = \\ = \mathsf{E}\left[\mathsf{E}\left[\mathbb{1}_{(U_i=g)}|\boldsymbol{\theta},\mathbb{Y}_i;\mathcal{C}_i\right]\middle|\mathbb{Y};\mathcal{C}\right] = \mathsf{E}\left[\mathbb{1}_{(U_i=g)}\middle|\mathbb{Y};\mathcal{C}\right],$$

we can use the sampled U_i^m for estimating the posterior mean of the classification probabilities. Specifically, it will be estimated by $\widehat{u}_{i,g} = \frac{1}{M}\sum\limits_{m=1}^{M}\mathbb{1}_{(U_i^m=g)}$, which is used for a crude classification by (P1) or (P2), see Section 3.2.1. Yet, only a point estimate for the posterior mean is obtained. To obtain some credible intervals (ET or HPD) one has to explore the full posterior distribution of classification probabilities.

This trick with sampled U_i^m also cannot be used when the classification probabilities for a newly observed unit not included in the training dataset are desired.

Deviance

Deviance is yet another general goodness-of-fit measure derived from the log-likelihood function. For given number of mixture components G, it is defined as

$$D^G\left(\boldsymbol{\theta};\mathbb{Y},\mathcal{C}\right) := -2\log p(\mathbb{Y}|\boldsymbol{\theta};\mathcal{C}) = -2\sum_{i=1}^{n}\log\left[\sum_{g=1}^{G} w_g p(\mathbb{Y}_i|U_i = g,\boldsymbol{\theta};\mathcal{C}_i)\right]. \quad (4.17)$$

In fact, this is the penalization-free baseline of AIC or BIC, for which one has to add $2d$ or $d\log n$ where d is the total number of unknown parameters. Aitkin et al. (2009) propose to decide about the two values $G_1 < G_2$ of the number of groups on the basis of the posterior probability

$$\mathsf{P}\left[D^{G_1}\left(\boldsymbol{\theta};\mathbb{Y},\mathcal{C}\right) > D^{G_2}\left(\boldsymbol{\theta};\mathbb{Y},\mathcal{C}\right)\middle|\mathbb{Y};\mathcal{C}\right]$$

that compares the deviances of the two nested models.

Yet, with more complicated data structures, which require presence of latent data $\mathcal{L}_i$, it is computationally demanding to evaluate the deviance. The contribution of a single unit i to the deviance has the form of

$$D_i^G = -2\ell(\mathbb{Y}_i|\boldsymbol{\theta};\mathcal{C}_i) = -2\log\left[\sum_{g=1}^{G} w_g\int p\left(\mathbb{Y}_i,\mathcal{L}_i|U_i = g,\boldsymbol{\theta};\mathcal{C}_i\right)\,\mathrm{d}\mathcal{L}_i\right], \quad (4.18)$$

where we recognize the summands from classification probabilities (3.4). Therefore, one has to evaluate $u_{i,g}\left(\boldsymbol{\theta}^m\right)$ for M values of $\boldsymbol{\theta}$, for each $g = 1,\dots,G$ and for each unit $i = 1,\dots,n$ to obtain the posterior distribution of D^G, which in total requires $M\cdot G\cdot n$ numerical approximations of the integral with respect to $\mathcal{L}_i$. Moreover, this process has to be repeated for $G\in\{1,\dots,G_{\max}\}$. Hence, efficient methods for approximation of the integral are desired.

Nevertheless, there are other ways on how to estimate the appropriate number of clusters within the Bayesian scope. One could simply assume a prior distribution for G and estimate it as an unknown parameter. However, when G is changed, so is the dimension of the parametric space. One has to construct a MCMC sampler capable of transition between parametric spaces of different dimensions, for example, using the *reversible jump* methodology. Another possible way, which will be taken for the *GLMM-based* model, is to set up the *sparse finite mixture* and let the MCMC sampler empty the redundant clusters, see Section 6.3.

4.4 Model hierarchy

We take a moment here to summarize and visualize the hierarchical structure of the two proposed models. They both assume a certain order in which elements of the model are created. In general, hyperparameters are set first. Then, randomized hyperparameters and parameters of interest are created. Unobserved latent data follow immediately afterwards. Only when all parameters and auxiliary variables are at disposal, the observed outcomes are generated.

This process of generating can be depicted in an oriented graph. It serves not only as an overview, but also helps to understand which elements truly come into play in full-conditional distributions. Moreover, the diagrams (Figures 4.3 and 4.4) also help to realize the similarities and the differences between the two considered models.

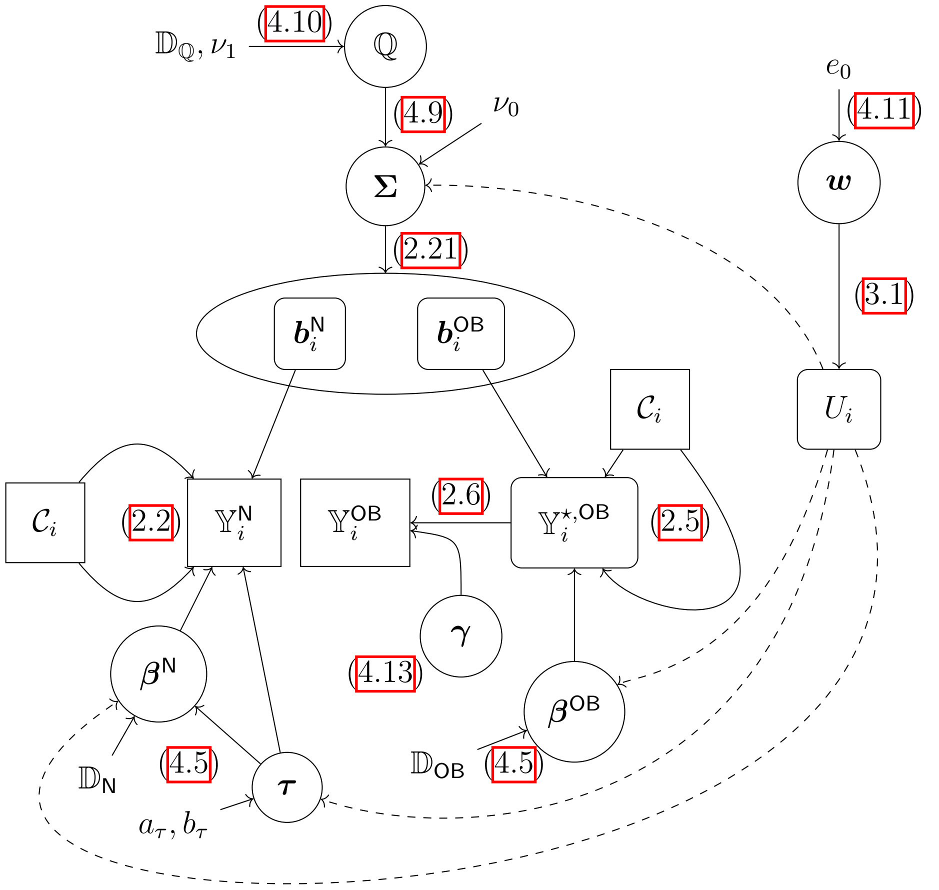
4.4.1 Hierarchy of the threshold concept model

The *threshold concept* model supposes that the observed binary and ordinal categories are secretly determined by the intervals, into which corresponding latent numeric outcome values belong to. Observed numeric outcomes and the latent numeric outcomes are modelled jointly through joint distribution of random effects with general covariance matrices $\boldsymbol{\Sigma}$, for which a structured prior is assumed. Simple Dirichlet distribution is assumed for marginal cluster probabilities $\boldsymbol{w}$ with fixed hyperparameter e_0.

To obtain a version of (4.16) tailored for this model, we have to combine the integrand of (3.5) with the prior distribution (4.3):

$$
\begin{aligned}
&p\left(\boldsymbol{w},\, \boldsymbol{\beta},\, \boldsymbol{\tau},\, \boldsymbol{\gamma},\, \boldsymbol{\Sigma},\, \boldsymbol{U},\, \boldsymbol{b},\, \mathbb{Y}^{\star,\mathsf{OB}},\, \mathbb{Q} \,\middle|\, \mathbb{Y}; \mathcal{C}\right) \propto \\
&\propto \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{OB}} \middle| \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{\gamma}\right) \cdot p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \middle| \boldsymbol{b}_i, \boldsymbol{\beta}^{(U_i)}, \boldsymbol{\tau}^{(U_i)}; \mathcal{C}_i\right) \cdot p\left(\boldsymbol{b}_i \middle| \boldsymbol{\Sigma}^{(U_i)}\right) \cdot p(U_i|\boldsymbol{w}) \cdot \\
&\qquad \cdot p(\boldsymbol{w}) \cdot p(\boldsymbol{\gamma}) \cdot p(\boldsymbol{\beta} \,|\, \boldsymbol{\tau}) \cdot p(\boldsymbol{\tau}) \cdot p(\boldsymbol{\Sigma}|\mathbb{Q}) \cdot p(\mathbb{Q}). \qquad (4.19)
\end{aligned}
$$

The whole model is summarized by the diagram in Figure 4.3.



$\mathbb{D}_{\mathbb{Q}}, \nu_1$
(4.10)
$\mathbb{Q}$
e_0
(4.9)
ν_0
(4.11)
$\boldsymbol{\Sigma}$
$\boldsymbol{w}$
(2.21)
(3.1)
$\boldsymbol{b}_i^{\mathsf{N}}$
$\boldsymbol{b}_i^{\mathsf{OB}}$
$\mathcal{C}_i$
U_i
$\mathcal{C}_i$
(2.2)
$\mathbb{Y}_i^{\mathsf{N}}$
$\mathbb{Y}_i^{\mathsf{OB}}$
(2.6)
$\mathbb{Y}_i^{\star,\mathsf{OB}}$
(2.5)
$\boldsymbol{\gamma}$
(4.13)
$\boldsymbol{\beta}^{\mathsf{N}}$
$\boldsymbol{\beta}^{\mathsf{OB}}$
$\mathbb{D}_{\mathsf{N}}$
(4.5)
$\boldsymbol{\tau}$
$\mathbb{D}_{\mathsf{OB}}$
(4.5)
a_τ, b_τ

Figure 4.3: Diagram depicting the hierarchy of the *threshold concept* model. Observed data (rectangles), latent unobserved data (rounded corners), parameters (circles), hyperparameters (not enclosed). Dashed line symbolizes a choice of the parameter depending on the current cluster.

4.4.2 Hierarchy of the GLMM-based model

The *GLMM-based* model structurally differs from the *threshold concept* model only in few aspects. There are no latent outcome values, each of the outcomes (possibly of 5 different types) is modelled directly by a certain GLMM. Instead of ordered thresholds $\boldsymbol{\gamma}$ we have ordered intercepts $\boldsymbol{c}$. Moreover, we allow for sparse finite mixture by controlling the hyperparamters for e_0, which sets up the prior of the marginal group allocation probabilities $\boldsymbol{w}$. The models for different outcomes are joined the same way in both models (through $\boldsymbol{b}_i$).

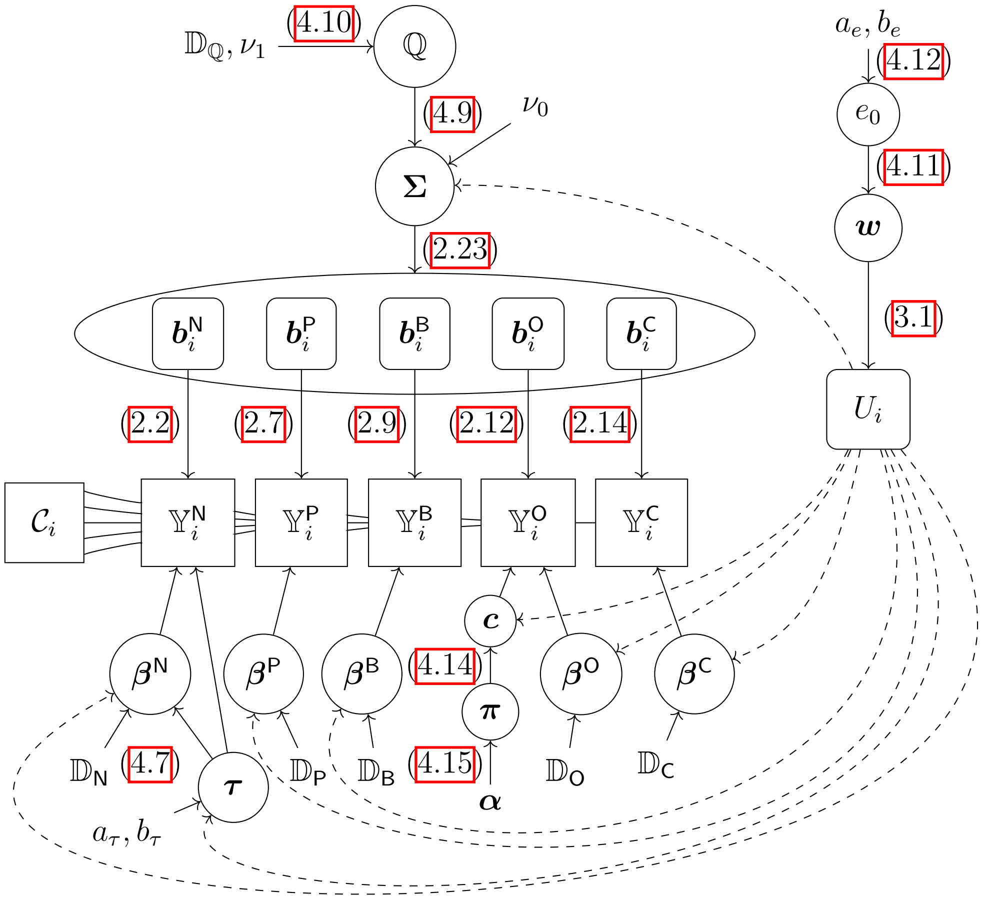
To obtain a version of (4.16) tailored for the *GLMM-based* model, we have to combine the integrand of (3.7) with the prior distribution (4.4):

$$p\left(\boldsymbol{w},\, \boldsymbol{\beta},\, \boldsymbol{\tau},\, \boldsymbol{c},\, \boldsymbol{\Sigma},\, \boldsymbol{U},\, \boldsymbol{b},\, \mathbb{Q},\, e_0 \,\middle|\, \mathbb{Y};\, \mathcal{C}\right) \propto$$

$$\propto \prod_{i=1}^{n} \left[\prod_{r\in\mathcal{R}} \prod_{j=1}^{n_i} p_{\mathsf{t}(r)} \left(Y_{i,j}^{r} \,\middle|\, \boldsymbol{b}_i^{r}, \boldsymbol{\beta}_r^{(U_i)}, \tau_r^{(U_i)}, \boldsymbol{c}_r^{(U_i)}; \mathcal{C}_{i,j} \right) \cdot p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\Sigma}^{(U_i)} \right) \cdot p(U_i|\boldsymbol{w}) \right] \cdot$$

$$\cdot\, p(\boldsymbol{w}|e_0) \cdot p(e_0) \cdot p(\boldsymbol{c}) \cdot p(\boldsymbol{\beta} \,|\, \boldsymbol{\tau}) \cdot p(\boldsymbol{\tau}) \cdot p(\boldsymbol{\Sigma}|\mathbb{Q}) \cdot p(\mathbb{Q}). \quad (4.20)$$

The whole model is summarized by the diagram in Figure 4.4.



$\mathbb{D}_{\mathbb{Q}}, \nu_1$
(4.10)
$\mathbb{Q}$
(4.9)
ν_0
$\boldsymbol{\Sigma}$
(2.23)
a_e, b_e
(4.12)
e_0
(4.11)
$\boldsymbol{w}$
(3.1)
U_i
$\boldsymbol{b}_i^{\mathsf{N}}$
$\boldsymbol{b}_i^{\mathsf{P}}$
$\boldsymbol{b}_i^{\mathsf{B}}$
$\boldsymbol{b}_i^{\mathsf{O}}$
$\boldsymbol{b}_i^{\mathsf{C}}$
(2.2)
(2.7)
(2.9)
(2.12)
(2.14)
$\mathcal{C}_i$
$\mathbb{Y}_i^{\mathsf{N}}$
$\mathbb{Y}_i^{\mathsf{P}}$
$\mathbb{Y}_i^{\mathsf{B}}$
$\mathbb{Y}_i^{\mathsf{O}}$
$\mathbb{Y}_i^{\mathsf{C}}$
$\boldsymbol{c}$
(4.14)
$\boldsymbol{\pi}$
(4.15)
$\boldsymbol{\alpha}$
$\boldsymbol{\beta}^{\mathsf{N}}$
$\boldsymbol{\beta}^{\mathsf{P}}$
$\boldsymbol{\beta}^{\mathsf{B}}$
$\boldsymbol{\beta}^{\mathsf{O}}$
$\boldsymbol{\beta}^{\mathsf{C}}$
$\mathbb{D}_{\mathsf{N}}$
(4.7)
$\boldsymbol{\tau}$
a_τ, b_τ
$\mathbb{D}_{\mathsf{P}}$
$\mathbb{D}_{\mathsf{B}}$
$\mathbb{D}_{\mathsf{O}}$
$\mathbb{D}_{\mathsf{C}}$

Figure 4.4: Diagram depicting the hierarchy of the *GLMM-based* model. Observed data (rectangles), latent unobserved data (rounded corners), parameters (circles), hyperparameters (not enclosed). Dashed line symbolizes a choice of the parameter depending on the current cluster.

5. MCMC estimation of the threshold concept model

This chapter is dedicated solely to the *threshold concept* model (Sections 2.4.1, 3.3.1, 4.4.1) and derivation of all necessary quantities to construct an MCMC algorithm. We list down all the full-conditional distributions which are all well known. Hence, Gibbs sampling is adopted without need for any Metropolis proposal steps. We discuss some of its strengths and weaknesses and then test its ability to estimate unknown parameters is a simulation study.

5.1 Full-conditional distributions

For this model, the set of all randomized elements (including the latent variables) of the model $\boldsymbol{\Psi}$ consists of $\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\tau}, \gamma, \boldsymbol{\Sigma}, \boldsymbol{U}, \boldsymbol{b}, \mathbb{Y}^{\star,\mathsf{OB}}$ and $\mathbb{Q}$. To derive full-conditional distributions for all parameters $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, we have to view the right hand side of (4.19) as a function of parameter $\boldsymbol{\psi}$, which can be decomposed into the following products:

$$\begin{aligned} &p\,(\boldsymbol{\psi}\,|\,\mathbb{Y},\,\boldsymbol{\Psi}_{-\boldsymbol{\psi}},\,\mathcal{H}_0;\,\mathcal{C}\,) \propto \\ &\propto \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{OB}}\middle|\mathbb{Y}_i^{\star,\mathsf{OB}},\gamma\right)\cdot p\left(\mathbb{Y}_i^{\mathsf{N}},\mathbb{Y}_i^{\star,\mathsf{OB}}\middle|\boldsymbol{b}_i,\boldsymbol{\beta}^{(U_i)},\boldsymbol{\tau}^{(U_i)};\mathcal{C}_i\right)\cdot p\left(\boldsymbol{b}_i\middle|\boldsymbol{\Sigma}^{(U_i)}\right)\cdot p(U_i|\boldsymbol{w})\cdot \\ &\quad\cdot p(\boldsymbol{w}|e_0)\cdot p(\gamma)\cdot p(\boldsymbol{\beta}\,|\,\boldsymbol{\tau},\boldsymbol{\beta}_0,\mathbb{D})\cdot p(\boldsymbol{\tau}|a_\tau,b_\tau)\cdot p(\boldsymbol{\Sigma}|\mathbb{Q},\nu_0)\cdot p\,(\mathbb{Q}\,|\mathbb{D}_{\mathbb{Q}},\nu_1\,)\,, \end{aligned} \tag{5.1}$$

where $\mathcal{H}_0$ denotes all fixed hyperparameters of prior distributions. Here, the last known values of latent quantities are used and not integrated as in (3.5), which considerably simplifies the evaluation. Since we work under proportionality, only factors including the current parameter of interest $\boldsymbol{\psi}$ remain and the rest is considered constant. One can equivalently read it from Figure 4.3, where full-conditional distribution of $\boldsymbol{\psi}$ consists of all arrows coming in and out of the corresponding node.

All derivations are made under the assumption that parameters $\boldsymbol{\beta}$, $\boldsymbol{\tau}$ and $\boldsymbol{\Sigma}$ are all group-specific. Similar derivations (with corresponding changes) can be made even under different setting of group-specificity of the parameters. Note that if $\boldsymbol{\tau}$ is group-specific, then $\boldsymbol{\beta}$ (at least the part corresponding to numeric outcomes) must also be group-specific to preserve the hierarchical structure of the prior distribution.

5.1.1 Cluster allocation probabilities $\boldsymbol{w}$

Prior probabilities $\boldsymbol{w}$ of belonging to a certain cluster, i.e. $w_g = \mathsf{P}\,[U_i = g]$, appear only in $p(U_i|\boldsymbol{w})$ and its general Dirichlet prior distribution $\mathsf{Dir}_G\,(e_1,\ldots,e_G)$ (in practice we use $e_g = e_0$ for all $g = 1,\ldots,G$). Therefore,

$$p\,(\boldsymbol{w}\,|\,\mathbb{Y},\,\boldsymbol{\Psi}_{-w},\,\mathcal{H}_0;\,\mathcal{C}\,) \propto \prod_{i=1}^{n} p\,(U_i|\boldsymbol{w})\cdot p(\boldsymbol{w}|e_1,\ldots,e_G)$$

can be simplified to

$$p\left(\boldsymbol{w}\,|\,\boldsymbol{U}, e_1, \ldots, e_G\right) \propto \prod_{i=1}^{n}\prod_{g=1}^{G} w_g^{\mathbb{1}_{(U_i=g)}} \cdot \prod_{g=1}^{G} w_g^{e_g-1} = \prod_{g=1}^{G} w_g^{n^{(g)}(\boldsymbol{U})+e_g-1},$$

where $n^{(g)}(\boldsymbol{U}) = \sum_{i=1}^{n} \mathbb{1}_{(U_i=g)}$ is the total number of units (from n possible) currently within the cluster g. We recognize the proportional shape of pdf of Dirichlet distribution, thus, under $e_g = e_0$ we have

$$\boldsymbol{w}\,\Big|\,\boldsymbol{U}, e_0 \sim \mathsf{Dir}_G\left(\boldsymbol{n}(\boldsymbol{U}) + e_0\boldsymbol{1}\right), \tag{5.2}$$

where $\boldsymbol{n}(\boldsymbol{U}) = \left(n^{(1)}(\boldsymbol{U}), \ldots, n^{(G)}(\boldsymbol{U})\right)^\top$.

5.1.2 Group-allocation indicators U_i

According to (5.1), the group-allocation indicator U_i of unit i appears in its prior distribution $U_i|\boldsymbol{w}$ and at places, where it selects the corresponding group-specific parameter:

$$p\left(U_i\,|\,\mathbb{Y}, \boldsymbol{\Psi}_{-U_i}, \mathcal{H}_0; \mathcal{C}\right) \propto p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}}\,\Big|\,\boldsymbol{b}_i, \boldsymbol{\beta}^{(U_i)}, \boldsymbol{\tau}^{(U_i)}; \mathcal{C}_i\right) \cdot p\left(\boldsymbol{b}_i\,\Big|\,\boldsymbol{\Sigma}^{(U_i)}\right) \cdot p\left(U_i|\boldsymbol{w}\right).$$

U_i only attains values $g \in \{1, \ldots, G\}$. Therefore, we aim to calculate the full-conditional probability that unit i is allocated in the group g:

$$\begin{aligned}\mathsf{P}\left[U_i = g\,\Big|\,\cdots\right] \propto\; & w_g \cdot \prod_{r\in\mathcal{R}^{\mathsf{Num}}} \left(\tau_r^{(g)}\right)^{\frac{n_i}{2}} \cdot \left|\boldsymbol{\Sigma}^{(g)}\right|^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-(g)}\boldsymbol{b}_i\right\} \cdot \\ & \cdot \exp\left\{-\frac{1}{2}\sum_{r\in\mathcal{R}^{\mathsf{Num}}}\sum_{j=1}^{n_i} \tau_r^{(g)}\left(Y_{i,j}^r - \eta_{i,j}^{r,(g)}\right)^2 - \frac{1}{2}\sum_{r\in\mathcal{R}^{\mathsf{OB}}}\sum_{j=1}^{n_i}\left(Y_{i,j}^{\star,r} - \eta_{i,j}^{r,(g)}\right)^2\right\},\end{aligned} \tag{5.3}$$

where $\eta_{i,j}^{r,(g)} = \left(\boldsymbol{x}_{i,j}^r\right)^\top \boldsymbol{\beta}_r^{(g)} + \left(\boldsymbol{z}_{i,j}^r\right)^\top \boldsymbol{b}_i^r$ is the linear predictor of j-th observation of outcome $r \in \mathcal{R}$ of unit i when belonging to the group g. Unlike the clustering probabilities (3.6), there is no integration involved since latent variables are at our disposal.

In case some parameter is not group-specific, the factors not depending on g could be skipped in the evaluation. The right hand side of (5.3) is first computed on log-scale, then a suitable constant is added to all G expressions to obtain reasonable exponentials which are then summed and proportionally compared to obtain the final probabilities.

5.1.3 Latent numeric variables $\mathbb{Y}^{\star,\mathsf{OB}}$

Latent numeric outcomes $\mathbb{Y}^{\star,\mathsf{OB}}$ for actually measured ordinal and binary outcomes $\mathbb{Y}^{\mathsf{OB}}$ appear only in the thresholding procedure (2.19) and the multivariate LME (2.5) for $\mathbb{Y}^{\star,\mathsf{OB}}$:

$$p\left(\mathbb{Y}^{\star,\mathsf{OB}}\,\Big|\,\mathbb{Y}, \boldsymbol{\Psi}_{-\mathbb{Y}^{\star,\mathsf{OB}}}; \mathcal{C}\right) \propto p\left(\mathbb{Y}^{\mathsf{OB}}\,\Big|\,\mathbb{Y}^{\star,\mathsf{OB}}, \boldsymbol{\gamma}\right) \cdot p\left(\mathbb{Y}^{\star,\mathsf{OB}}\,\Big|\,\boldsymbol{b}, \boldsymbol{U}, \boldsymbol{\beta}, \boldsymbol{\tau}; \mathcal{C}\right).$$

Given random effects, both thresholding and LME for the latent variables are independent for all $r \in \mathcal{R}^{\mathsf{OB}}$, $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Ignoring the thresholding, $Y_{i,j}^{\star,r}$ would follow $\mathsf{N}\left(\eta_{i,j}^{(U_i),r}, 1\right)$ according to (2.5), however, corresponding density is now limited by indicator $\mathbb{1}_{\left(\gamma^r_{k-1}\,,\,\gamma^r_k\right]}\left(Y^{\star,r}_{i,j}\right)$ from (2.19), where $k = Y^r_{i,j}$. Therefore, the full-conditional distribution is the truncated normal distribution on the interval $\left(\gamma^r_{k-1}\,,\,\gamma^r_k\right]$:

$$Y^{\star,r}_{i,j}\,\Big|\,Y^r_{i,j} = k, \boldsymbol{b}_i, U_i, \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathcal{C} \sim \mathsf{TN}\left(\eta^{(U_i),r}_{i,j},\, 1,\, \gamma^r_{k-1},\, \gamma^r_k\right). \tag{5.4}$$

5.1.4 Thresholds $\boldsymbol{\gamma}$

Parameter $\boldsymbol{\gamma}$ influences (5.1) only in the thresholding phase (2.19) and in the prior distribution (4.13) of $\boldsymbol{\gamma}$:

$$p\left(\boldsymbol{\gamma}\,|\,\mathbb{Y}, \boldsymbol{\Psi}_{-\gamma^r}; \mathcal{C}\right) \propto \prod_{i=1}^{n} p\left(\mathbb{Y}^{\mathsf{OB}}_i\,\Big|\,\mathbb{Y}^{\star,\mathsf{OB}}_i, \boldsymbol{\gamma}\right)\cdot p\left(\boldsymbol{\gamma}\right).$$

Let us consider ordinal outcome $r \in \mathcal{R}^{\mathsf{Ord}}$ and the corresponding set of thresholds: $-\infty = \gamma^r_{-1}, \gamma^r_0, \boldsymbol{\gamma}^r, \gamma^r_{K^r-1} = \infty$. Let $\mathcal{Y}^r_k$ be the set of all latent numeric outcomes $Y^{\star,r}_{i,j}$ such that the truly measured ordinal category is $k = 0, \dots, K^r - 1$, i.e.

$$\mathcal{Y}^r_k = \left\{Y^{\star,r}_{i,j} : Y^r_{i,j} = k, \quad i = 1, \dots, n, \quad j = 1, \dots, n_i\right\},$$

which is assumed to be non-empty (all levels of outcome K^r are attained at least once). The latent numeric variables had to be generated according to the threshold concept, therefore, the following inequalities hold:

$$-\infty < \underset{\in\mathcal{Y}^r_0}{y_0} < \gamma^r_0 < \underset{\in\mathcal{Y}^r_1}{y_1} < \gamma^r_1 < \underset{\in\mathcal{Y}^r_2}{y_2} < \dots < \gamma^r_{K^r-2} < \underset{\in\mathcal{Y}^r_{K^r-1}}{y_{K^r-1}} < \infty.$$

Had $\boldsymbol{\gamma}$ been allowed to be group-specific parameter, these inequalities could have been broken by units switching their current allocation. Under the uniform prior (4.13) for $\boldsymbol{\gamma}^r$ we get that the individual thresholds γ^r_k are uniformly distributed on intervals given by maxima and minima of the corresponding sets:

$$\gamma^r_k\,|\,\boldsymbol{Y}^r, \boldsymbol{Y}^{\star,r} \;\sim\; \mathsf{Unif}\left[\max_{y\in\mathcal{Y}^r_k} y,\; \min_{y\in\mathcal{Y}^r_{k+1}} y\right], \qquad k = 1, \dots, K^r - 2. \tag{5.5}$$

5.1.5 Precision parameters $\boldsymbol{\tau}$

Parameters $\boldsymbol{\tau} = \left\{\tau^{(g)}_r :\; g = 1, \dots, G,\; r \in \mathcal{R}^{\mathsf{Num}}\right\}$ are the inverse variance of errors of the supposed LME models over numeric outcomes. The right-hand side of (5.1) includes $\boldsymbol{\tau}$ in the supposed LME for $\mathbb{Y}^{\mathsf{N}}_i$ and the prior distribution of $(\boldsymbol{\beta}, \boldsymbol{\tau})$:

$$p\left(\boldsymbol{\tau}\,|\,\mathbb{Y}, \boldsymbol{\Psi}_{-\boldsymbol{\tau}}, \mathcal{H}_0; \mathcal{C}\right) \propto \prod_{i=1}^{n} p\left(\mathbb{Y}^{\mathsf{N}}_i\,\Big|\,\boldsymbol{b}_i, \boldsymbol{\beta}^{(U_i)}, \boldsymbol{\tau}^{(U_i)}; \mathcal{C}_i\right)\cdot p\left(\boldsymbol{\beta}|\boldsymbol{\tau}; \boldsymbol{\beta}_0, \mathbb{D}\right)\cdot p\left(\boldsymbol{\tau}|a_\tau, b_\tau\right).$$

From the product structure of (2.20) and prior (4.5), we see that individual $\tau^{(g)}_r$ are distributed independently of each other given other parameters. Then, the pdf of the full-conditional distribution of single $\tau^{(g)}_r$ is proportional to

$$p\left(\tau^{(g)}_r\,\Big|\,\boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{\beta}^{(g)}_r, \boldsymbol{\beta}_{0,r}, \mathbb{D}_r, a_\tau, b_\tau; \mathcal{C}\right) \propto \left(\tau^{(g)}_r\right)^{\frac{1}{2}\sum\limits_{i\in\mathcal{N}_g(\boldsymbol{U})} n_i \;+\frac{1}{2}d^{\mathsf{F}}_r + a_\tau - 1}\cdot$$
$$\cdot\exp\left\{-\tau^{(g)}_r\left[\frac{1}{2}\sum_{i\in\mathcal{N}_g(\boldsymbol{U})}\sum_{j=1}^{n_i}\left(Y^r_{i,j} - \eta^{r,(g)}_{i,j}\right)^2 + \frac{1}{2}\sum_{j=1}^{d^{\mathsf{F}}_r}\frac{\left(\beta^{(g)}_{r,j} - \beta^r_{0,r,j}\right)^2}{d^r_{j,j}} + b_\tau\right]\right\},$$

where $\mathcal{N}_g(\boldsymbol{U}) = \{i : U_i = g,\ i = 1, \dots, n\}$ is a set of units currently belonging to group g. For $\boldsymbol{Y}^r, \mathcal{C}$ and current values of $\boldsymbol{U}, \boldsymbol{b}^r$ and $\boldsymbol{\beta}_r^{(g)}$ let us denote

$$\widetilde{a}_{\tau,r}^{(g)} = \frac{1}{2} \sum_{i \in \mathcal{N}_g(\boldsymbol{U})} n_i \;+ \frac{d_r^{\mathsf{F}}}{2} + a_\tau,$$

$$\widetilde{b}_{\tau,r}^{(g)} = \frac{1}{2} \sum_{i \in \mathcal{N}_g(\boldsymbol{U})} \sum_{j=1}^{n_i} \left(Y_{i,j}^r - \eta_{i,j}^{r,(g)}\right)^2 + \frac{1}{2}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right) + b_\tau.$$

Under this notation, we see that

$$\tau_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{\beta}_r^{(k)}, \boldsymbol{\beta}_{0,r}, \mathbb{D}_r, a_\tau, b_\tau; \mathcal{C} \;\sim\; \Gamma\left(\widetilde{a}_{\tau,r}^{(g)}, \widetilde{b}_{\tau,r}^{(g)}\right) \tag{5.6}$$

independently for each $r \in \mathcal{R}^{\mathsf{Num}}$ and $g = 1, \dots, G$.

5.1.6 Fixed effects $\boldsymbol{\beta}$

Fixed effects $\boldsymbol{\beta}$ appear in (5.1) only in the LME model specification (2.1), (2.5) and prior distribution (4.5):

$$p\left(\boldsymbol{\beta} \,|\, \mathbb{Y}, \boldsymbol{\Psi}_{-\beta}, \mathcal{H}_0; \mathcal{C}\right) \propto \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \,\middle|\, \boldsymbol{b}_i, \boldsymbol{\beta}^{(U_i)}, \boldsymbol{\tau}^{(U_i)}; \mathcal{C}_i\right) \cdot p\left(\boldsymbol{\beta} | \boldsymbol{\tau}, \boldsymbol{\beta}_0, \mathbb{D}\right),$$

which can be decomposed for outcomes $r \in \mathcal{R}$ and $g = 1, \dots, G$ as follows:

$$p\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \cdots\right) \propto \exp\left\{-\frac{\tau_r^{(g)}}{2}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)\right\} \cdot$$
$$\cdot \exp\left\{-\frac{\tau_r^{(g)}}{2}\left(\widetilde{\boldsymbol{Y}}_{\mathcal{N}_g(\boldsymbol{U})}^r - \mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r \boldsymbol{\beta}_r^{(g)}\right)^\top \left(\widetilde{\boldsymbol{Y}}_{\mathcal{N}_g(\boldsymbol{U})}^r - \mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r \boldsymbol{\beta}_r^{(g)}\right)\right\},$$

where notation $\bullet_{\mathcal{N}_g(\boldsymbol{U})}$ restricts the expression $\bullet$ to the subset of units in group g:

$$\mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r = \begin{pmatrix} \vdots \\ \mathbb{X}_i^r \\ \vdots \end{pmatrix}, i \in \mathcal{N}_g(\boldsymbol{U}),$$

$$\widetilde{\boldsymbol{Y}}_{\mathcal{N}_g(\boldsymbol{U})}^r = \begin{cases} \left\{\left(\boldsymbol{Y}_i^r - \mathbb{Z}_i^r \boldsymbol{b}_i^r\right)^\top, i \in \mathcal{N}_g(\boldsymbol{U})\right\}, & \text{if } r \in \mathcal{R}^{\mathsf{Num}}, \\ \left\{\left(\boldsymbol{Y}_i^{\star,r} - \mathbb{Z}_i^r \boldsymbol{b}_i^r\right)^\top, i \in \mathcal{N}_g(\boldsymbol{U})\right\}, & \text{if } r \in \mathcal{R}^{\mathsf{OB}}. \end{cases}$$

Using basic algebraic operations and ignoring several multiplicative constants, we can rewrite the pdf of full-conditional distribution of $\boldsymbol{\beta}_r^{(g)}$ into:

$$p\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \tau_r^{(g)}, \boldsymbol{\beta}_{0,r}, \mathbb{D}_r; \mathcal{C}\right) \propto$$
$$\exp\left\{-\frac{\tau_r^{(g)}}{2}\left(\boldsymbol{\beta}_r^{(g)} - \widetilde{\boldsymbol{\beta}}_r^{(g)}\right)^\top \left[\left(\mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r\right)^\top \mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r + \mathbb{D}_r^{-1}\right]\left(\boldsymbol{\beta}_r^{(g)} - \widetilde{\boldsymbol{\beta}}_r^{(g)}\right)\right\},$$

where

$$\widetilde{\boldsymbol{\beta}}_r^{(g)} = \left[\left(\mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r\right)^\top \mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r + \mathbb{D}_r^{-1}\right]^{-1}\left(\left(\mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r\right)^\top \widetilde{\boldsymbol{Y}}_{\mathcal{N}_g(\boldsymbol{U})}^r + \mathbb{D}_r^{-1}\boldsymbol{\beta}_{0,r}\right),$$

which compared to the pdf of multivariate normal distribution yields

$$\boldsymbol{\beta}_r^{(g)} \,\Big|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \tau_r^{(g)}, \boldsymbol{\beta}_{0,r}, \mathbb{D}_r; \mathcal{C} \;\sim\; \mathsf{N}_{d_r^{\mathsf{F}}}\left(\widetilde{\boldsymbol{\beta}}_r^{(g)},\, \frac{1}{\tau_r^{(g)}}\left[\left(\mathbb{X}^r_{\mathcal{N}_g(\boldsymbol{U})}\right)^{\top} \mathbb{X}^r_{\mathcal{N}_g(\boldsymbol{U})} + \mathbb{D}_r^{-1}\right]^{-1}\right). \tag{5.7}$$

The variance matrix is symmetric and positive-definite (since $A^{\top}A \geq 0$ in general for any matrix A and $\mathbb{D}_r^{-1} > 0$), hence, *Cholesky decomposition* (see Section 7.4, Algorithm 5) is in practice used to efficiently sample from this distribution. This is a typical example where the prior distribution regularizes the likelihood contribution.

5.1.7 Scale matrix $\mathbb{Q}$ for $\boldsymbol{\Sigma}$

$\mathbb{Q}$ is an auxiliary parameter that makes prior distribution of matrices $\boldsymbol{\Sigma}$ (4.9) more flexible within Gibbs sampler. The right-hand side of (5.1) shrinks into

$$p\,(\mathbb{Q} \,|\, \mathbb{Y}, \boldsymbol{\Psi}_{-\mathbb{Q}}, \mathcal{H}_0; \mathcal{C}\,) \propto p\,(\boldsymbol{\Sigma} \,|\, \mathbb{Q}, \nu_0) \cdot p\,(\mathbb{Q} \,|\, \mathbb{D}_{\mathbb{Q}}, \nu_1\,)\,,$$

where the combined pdfs on the right hand side correspond to Wishart distribution. Since $\boldsymbol{\Sigma}$ is considered group-specific, $\mathbb{Q}$ is assumed to give rise to G matrices $\boldsymbol{\Sigma}^{(g)}$. Combining the $G+1$ Wishart pdfs (4.9) and (4.10) we obtain

$$p\left(\mathbb{Q}^{-1} \,\Big|\, \boldsymbol{\Sigma}, \nu_0, \nu_1, \mathbb{D}_{\mathbb{Q}}\right) \propto \left|\mathbb{Q}^{-1}\right|^{\frac{G\nu_0+\nu_1-d^{\mathsf{R}}-1}{2}} \exp\left\{-\frac{1}{2}\,\mathsf{Tr}\left[\left(\sum_{g=1}^{G} \boldsymbol{\Sigma}^{-(g)} + \mathbb{D}_{\mathbb{Q}}^{-1}\right)\mathbb{Q}^{-1}\right]\right\},$$

which resembles again a pdf of Wishart distribution. Therefore,

$$\mathbb{Q}^{-1} \,\Big|\, \boldsymbol{\Sigma}, \nu_0, \nu_1, \mathbb{D}_{\mathbb{Q}} \;\sim\; \mathsf{W}_{d^{\mathsf{R}}}\left(\left[\sum_{g=1}^{G} \boldsymbol{\Sigma}^{-(g)} + \mathbb{D}_{\mathbb{Q}}^{-1}\right]^{-1}, G\nu_0 + \nu_1\right). \tag{5.8}$$

From a practical point of view, it is convenient to work directly with $\mathbb{Q}^{-1}$. The inversion to $\mathbb{Q}$ is unnecessary.

5.1.8 Covariance matrices $\boldsymbol{\Sigma}$ for random effects $\boldsymbol{b}$

Parameter $\boldsymbol{\Sigma}$ is the set of G covariance matrices $\boldsymbol{\Sigma}^{(g)}$ for random effects $\boldsymbol{b}_i$ that contributes to the right-hand side of (5.1) in the pdf for random effects (2.21) and in the prior distribution of $\boldsymbol{\Sigma}$ given by (4.9):

$$p\,(\boldsymbol{\Sigma} \,|\, \mathbb{Y}, \boldsymbol{\Psi}_{-\boldsymbol{\Sigma}}, \mathcal{H}_0; \mathcal{C}\,) \propto \prod_{i=1}^{n} p\left(\boldsymbol{b}_i \,\Big|\, \boldsymbol{\Sigma}^{(U_i)}\right) \cdot p\,(\boldsymbol{\Sigma} \,|\, \mathbb{Q}, \nu_0)\,.$$

Again, we need to divide units into the groups $\mathcal{N}_g(\boldsymbol{U}), g = 1, \ldots, G$ according to their current allocation indicators $\boldsymbol{U}$. The equation above decomposes into G independent parts – one for each group $g = 1, \ldots, G$. Considering the group g, the right-hand side of the equation above reduces into

$$\begin{aligned} p\left(\boldsymbol{\Sigma}^{-(g)} \,\Big|\, \boldsymbol{U}, \boldsymbol{b}, \mathbb{Q}, \nu_0\right) &\propto \\ &\propto \left|\boldsymbol{\Sigma}^{-(g)}\right|^{\frac{n^{(g)}(\boldsymbol{U})+\nu_0-d^{\mathsf{R}}-1}{2}} \exp\left\{-\frac{1}{2}\,\mathsf{Tr}\left[\left(\mathbb{Q}^{-1} + \sum_{i\in\mathcal{N}_g(\boldsymbol{U})} \boldsymbol{b}_i\boldsymbol{b}_i^{\top}\right)\boldsymbol{\Sigma}^{-(g)}\right]\right\}, \end{aligned}$$

which again resembles a pdf of Wishart distribution for the precision matrix $\boldsymbol{\Sigma}^{-(g)} = \left(\boldsymbol{\Sigma}^{(g)}\right)^{-1}$. Therefore, independently for all $g = 1, \ldots, G$

$$\boldsymbol{\Sigma}^{-(g)} \,\Big|\, \boldsymbol{U}, \boldsymbol{b}, \mathbb{Q}, \nu_0 \;\sim\; \mathsf{W}_{d^{\mathsf{R}}}\left(\widetilde{\mathbb{Q}}^{(g)}, n^{(g)}(\boldsymbol{U}) + \nu_0\right), \tag{5.9}$$

where

$$\widetilde{\mathbb{Q}}^{(g)} = \left(\widetilde{\mathbb{Q}}^{-(g)}\right)^{-1} \quad \text{and} \quad \widetilde{\mathbb{Q}}^{-(g)} = \mathbb{Q}^{-1} + \sum_{i \in \mathcal{N}_g(\boldsymbol{U})} \boldsymbol{b}_i \boldsymbol{b}_i^\top.$$

From a practical point of view, it is convenient to work directly with $\boldsymbol{\Sigma}^{-(g)}$. The inversion to $\boldsymbol{\Sigma}^{(g)}$ is redundant since sampling random effects $\boldsymbol{b}_i$ requires (as we will soon see) only the precision matrix $\boldsymbol{\Sigma}^{-(g)}$.

5.1.9 Random effects $\boldsymbol{b}$

The key role of our model is played by the random effects $\boldsymbol{b}_i$, $i = 1, \ldots, n$ that create linear predictors $\eta_{i,j}^{r,(g)}$, $g = 1, \ldots, G$, $r \in \mathcal{R}$ and $j = 1, \ldots, n_i$. The pdf of corresponding full-conditional distribution (5.1) is based on the assumed LME (2.1), (2.5) and the joint distribution of random effects (2.21):

$$p\left(\boldsymbol{b} \,|\, \mathbb{Y}, \boldsymbol{\Psi}_{-\boldsymbol{b}}, \mathcal{H}_0; \mathcal{C}\right) \propto \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \,\Big|\, \boldsymbol{b}_i, \boldsymbol{\beta}^{(U_i)}, \boldsymbol{\tau}^{(U_i)}; \mathcal{C}_i\right) \cdot \prod_{i=1}^{n} p\left(\boldsymbol{b}_i \,\Big|\, \boldsymbol{\Sigma}^{(U_i)}\right).$$

Clearly, random effects $\boldsymbol{b}_i$ will be distributed independently even in the full-conditional distribution.

Let us select a unit i (say from group $U_i = g$) in which case its corresponding pdf is of the shape

$$\begin{aligned} p\left(\boldsymbol{b}_i \,|\, \cdots\right) \propto & \prod_{r \in \mathcal{R}^{\mathsf{Num}}} \exp\left\{-\frac{\tau_r^{(g)}}{2}\left(\widetilde{\boldsymbol{Y}}_i^r - \mathbb{Z}_i^r \boldsymbol{b}_i^r\right)^\top \left(\widetilde{\boldsymbol{Y}}_i^r - \mathbb{Z}_i^r \boldsymbol{b}_i^r\right)\right\} \cdot \\ & \cdot \prod_{r \in \mathcal{R}^{\mathsf{OB}}} \exp\left\{-\frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_i^{\star,r} - \mathbb{Z}_i^r \boldsymbol{b}_i^r\right)^\top \left(\widetilde{\boldsymbol{Y}}_i^{\star,r} - \mathbb{Z}_i^r \boldsymbol{b}_i^r\right)\right\} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-(g)} \boldsymbol{b}_i\right\}, \end{aligned}$$

where $\widetilde{\boldsymbol{Y}}_i^r = \boldsymbol{Y}_i^r - \mathbb{X}_i^r \boldsymbol{\beta}_r^{(g)}$ and $\widetilde{\boldsymbol{Y}}_i^{\star,r} = \boldsymbol{Y}_i^{\star,r} - \mathbb{X}_i^r \boldsymbol{\beta}_r^{(g)}$. Constructing

$$\widetilde{\boldsymbol{Y}}_i = \begin{pmatrix} \vdots \\ \sqrt{\tau_r^{(g)}}\widetilde{\boldsymbol{Y}}_i^r \\ \vdots \\ \widetilde{\boldsymbol{Y}}_i^{\star,r} \\ \vdots \end{pmatrix} \begin{matrix} \\ r \in \mathcal{R}^{\mathsf{Num}}, \\ \\ r \in \mathcal{R}^{\mathsf{OB}}, \\ \\ \end{matrix}, \qquad \widetilde{\mathbb{Z}}_i = \begin{pmatrix} \ddots & & & \\ & \sqrt{\tau_r^{(g)}}\mathbb{Z}_i^r & & \\ & & \ddots & \\ & & & \mathbb{Z}_i^r & \\ & & & & \ddots \end{pmatrix},$$

we can simplify the above to

$$\exp\left\{-\frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{Z}}_i \boldsymbol{b}_i\right)^\top \left(\widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{Z}}_i \boldsymbol{b}_i\right) - \frac{1}{2}\boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-(g)} \boldsymbol{b}_i\right\},$$

which after several algebraic operations and ignoring multiplicative constants becomes

$$\exp\left\{-\frac{1}{2}\left(\widetilde{\boldsymbol{b}}_i - \boldsymbol{b}_i\right)^\top \left[\widetilde{\mathbb{Z}}_i^\top \widetilde{\mathbb{Z}}_i + \boldsymbol{\Sigma}^{-(g)}\right] \left(\widetilde{\boldsymbol{b}}_i - \boldsymbol{b}_i\right)\right\},$$

where $\widetilde{\boldsymbol{b}}_i = \left[\widetilde{\mathbb{Z}}_i^\top \widetilde{\mathbb{Z}}_i + \boldsymbol{\Sigma}^{-(g)}\right]^{-1} \left(\widetilde{\mathbb{Z}}_i^\top \widetilde{\boldsymbol{Y}_i}\right)$. Therefore, the full-conditional distribution of $\boldsymbol{b}_i$ for a unit belonging to group $g = 1, \dots, G$ is

$$\boldsymbol{b}_i \,\Big|\, \mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}}, U_i = g, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\Sigma}; \mathcal{C}_i \;\sim\; \mathsf{N}_{d^{\mathsf{R}}}\left(\widetilde{\boldsymbol{b}}_i,\, \left[\widetilde{\mathbb{Z}}_i^\top \widetilde{\mathbb{Z}}_i + \boldsymbol{\Sigma}^{-(g)}\right]^{-1}\right). \tag{5.10}$$

Similarly as for the fixed effects, utilizing Cholesky decomposition of the precision matrix is an efficient way for sampling from this distribution, see Section 7.4, Algorithm 5.

5.2 Gibbs sampling algorithm

Posterior distribution $p(\boldsymbol{\Psi}|\mathbb{Y};\mathcal{C})$ of $\boldsymbol{\Psi}$ (consisting of the model parameters $\boldsymbol{\theta}$, latent variables $\mathcal{L}$ and randomized hyperparameters $\mathcal{H}$) and any parametric function of $\boldsymbol{\Psi}$ will be estimated using MCMC methodology, see Section 4.3. We adopted the well known *Gibbs sampling* scheme in which a new value of each of the parameters is sampled from its full-conditional distribution while always utilizing the last known values of other parameters. Due to our (semi)-conjugate choice of prior distributions, the full-conditional distributions are from well-known families, hence, straightforward to be sampled from. In previous section, we derived the required distributions and now we summarize them into the Gibbs sampling algorithm, see Algorithm 1.

Let us discuss more the initialization phase of the algorithm. Once the data of longitudinal format are passed down, we have to set values of technical and prior distribution parameters. Especially, the number of clusters G, the length of the burn-in period B (initial thrown-away part of the chain), the desired length of the remaining chain M (multiplied by the required thinning parameter). It is always a good idea to sample several such chains in parallel to check convergence to the stationary distribution to evade local posterior modes. Each chain requires different initial values of the sampled states $\boldsymbol{\Psi}$, which could be easily achieved by generating random partitions of n units into G clusters and estimating the group-specific parameters from the data partitioned into the initial clusters.

The individual initialization steps given in Algorithm 1 could be replaced by simpler solutions. For example, sampling random effects $\boldsymbol{b}_i^0$ from $\mathsf{N}\left(0,\, \sigma^2\right)$ with fixed low σ^2 or even fixing $\boldsymbol{b}_i^0 = \mathbf{0}$ and $\boldsymbol{\Sigma}^{(g),0} = \mathsf{diag}(10^2)$. One can use Algorithm 1 with $B > 0$ and $M = 0$ to save the last known state $\boldsymbol{\Psi}^B$ and use it as initial values in the next iteration of the algorithm.

Though, sampling from the full-conditionals seems elegant at first sight, some steps slow down the convergence. For example, when started from inappropriate values of $\boldsymbol{\gamma}$, it takes even tens of thousands steps to converge. The length of the convergence phase heavily depends on the size of the dataset; with the `PBC910` of 918 datarows we barely notice any problems. However, with the EU-SILC database of $27\,386 \times 4$ rows we witness the extremely slow convergence. The problem comes from the structure of (5.5), where we sample uniformly from an interval given by the maximal latent outcome value of all rows with the observed $Y_{i,j}^r = k$ and the minimal latent outcome value of all rows with the observed $Y_{i,j}^r = k + 1$. Hence, with larger sample size the interval becomes naturally thinner. Therefore, the window of opportunity to change is negligible, hence the very slow convergence. Once the analyst becomes familiar with the dataset, the initial value $\boldsymbol{\gamma}^0$ can be manually adjusted to avoid long waiting times of reaching the stationary distribution. Slow convergence is observed also for other parameters tied with the latent quantities, e.g. $\boldsymbol{\Sigma}$, although several hundred of *burn-in* steps are enough to eliminate such problem in this case even for large datasets.

Algorithm 1 Gibbs sampling for the *threshold concept* model

```
Input: Data $\mathbb{Y}$ of longitudinal profiles of $n$ units and covariates $\mathcal{C}$.
Set the length of the burn-in period $B$ and the final length of the chains $M$.
Choose the number of clusters $G$ and set the fixed hyperparameters $\mathcal{H}_0$.
Declare or find initial values $\boldsymbol{\Psi}^0$ in this order:
    • divide $n$ units randomly into $G$ clusters by $U_i^0 \sim \mathsf{Unif}\{1, \ldots, G\}$;
    • estimate the thresholds $\boldsymbol{\gamma}^0$ to fit the proportions of $\mathbb{Y}^{\mathsf{O}}$ to $\mathsf{N}(0,\, 1)$;
    • take some values from intervals given by $\boldsymbol{\gamma}^0$ for latent outcomes $\mathbb{Y}^{\star,\mathsf{OB}}$;
    • estimate $\boldsymbol{\beta}^0$ using linear regression ignoring random effects;
    • estimate $\boldsymbol{b}_i^0$ using linear regression of (latent) outcomes lowered by the fixed part of the predictor;
    • estimate $\boldsymbol{\tau}^0$ by comparing the predictor to (latent) outcomes;
    • estimate $\boldsymbol{\Sigma}^0$ from $\boldsymbol{b}_i^0$ by a sample covariance matrix;
    • estimate $\mathbb{Q}^0$ by inverting the mean of all $\boldsymbol{\Sigma}^{(g),0}$ divided by $\nu_0$.

Gibbs sampling Always use the very last known values of other parameters,
    i.e. either from $\boldsymbol{\Psi}^{m-1}$ or $\boldsymbol{\Psi}^m$:
    for $m$ in $1 : (B + M)$ do
        • $\boldsymbol{w}^m \,|\, \boldsymbol{U}, e_0 \overset{(5.2)}{\sim} \mathsf{Dir}_G\left(\boldsymbol{n}(\boldsymbol{U}) + e_0\mathbf{1}\right)$;
        • $U_i^m \;\sim\; \mathsf{P}\left[U_i = g \,\middle|\, \mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{b}_i, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\Sigma}; \mathcal{C}_i\right] \overset{(5.3)}{=} \cdots$;
        • $Y_{i,j}^{\star,r,m} \,\middle|\, Y_{i,j}^r = k, \boldsymbol{b}_i, U_i, \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathcal{C} \overset{(5.4)}{\sim} \mathsf{TN}\left(\eta_{i,j}^{(U_i),r},\, 1,\, \gamma_{k-1}^r,\, \gamma_k^r\right)$   for $r \in \mathcal{R}^{\mathsf{OB}}$;
        • $\gamma_k^{r,m} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{Y}^{\star,r} \overset{(5.5)}{\sim} \mathsf{Unif}\left[\max_{y \in \mathcal{Y}_k^r} y,\, \min_{y \in \mathcal{Y}_{k+1}^r} y\right]$   for $r \in \mathcal{R}^{\mathsf{Ord}}$, $k = 1, \ldots, K^r - 2$;
        • $\tau_r^{(g),m} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{\beta}_r^{(g)}, \boldsymbol{\beta}_{0,r}, \mathbb{D}_r, a_\tau, b_\tau; \mathcal{C} \overset{(5.6)}{\sim} \Gamma\left(\widetilde{a}_{\tau,r}^{(g)}, \widetilde{b}_{\tau,r}^{(g)}\right)$   for $r \in \mathcal{R}^{\mathsf{Num}}$;
        • $\boldsymbol{\beta}_r^{(g),m} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \tau_r^{(g)}, \boldsymbol{\beta}_{0,r}, \mathbb{D}_r; \mathcal{C} \overset{(5.7)}{\sim}$
              $\mathsf{N}_{d_r^{\mathsf{F}}}\left(\widetilde{\boldsymbol{\beta}}_r^{(g)}, \frac{1}{\tau_r^{(g)}}\left[\left(\mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r\right)^\top \mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r + \mathbb{D}_r^{-1}\right]^{-1}\right)$   for $r \in \mathcal{R}$;
        • $(\mathbb{Q}^m)^{-1} \,\middle|\, \boldsymbol{\Sigma}, \nu_0, \nu_1, \mathbb{D}_{\mathbb{Q}} \overset{(5.8)}{\sim} \mathsf{W}_{d^{\mathsf{R}}}\left(\left[\sum_{g=1}^{G} \boldsymbol{\Sigma}^{-(g)} + \mathbb{D}_{\mathbb{Q}}^{-1}\right]^{-1}, G\nu_0 + \nu_1\right)$;
        • $\boldsymbol{\Sigma}^{-(g),m} \,|\, \boldsymbol{U}, \boldsymbol{b}, \mathbb{Q}, \nu_0 \overset{(5.9)}{\sim} \mathsf{W}_{d^{\mathsf{R}}}\left(\widetilde{\mathbb{Q}}^{(g)}, n^{(g)}(\boldsymbol{U}) + \nu_0\right)$;
        • $\boldsymbol{b}_i^m \,\middle|\, \mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}}, U_i = g, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\Sigma}; \mathcal{C}_i \overset{(5.10)}{\sim} \mathsf{N}_{d^{\mathsf{R}}}\left(\widetilde{\boldsymbol{b}}_i, \left[\widetilde{\mathbb{Z}}_i^\top \widetilde{\mathbb{Z}}_i + \boldsymbol{\Sigma}^{-(g)}\right]^{-1}\right)$.
    end for
```

5.3 Simulation study

To demonstrate the functionality of the implemented methodology for the *threshold concept* model, we performed a simulation study. To this end, data consisting of a numeric, a binary and an ordinal variable were generated under the assumption of different types of random effects structure. The only parameter distinguishing the latent groups ($G = 2$ or $G = 3$) was the parameter connected to the parametrization of time, i.e. intercept or slope. Parameters describing the covariance structure ($\boldsymbol{\tau}$ and $\boldsymbol{\Sigma}^{-1}$) were held equal for all latent groups.

Simulation design

Each type of response (numeric, ordinal and binary) is represented by only one longitudinally measured variable ($Y^{\mathsf{N}}_{i,j}$, $Y^{\mathsf{O}}_{i,j}$, $Y^{\mathsf{B}}_{i,j}$, $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$). We set the number n_i of observations per one unit to be fixed at $n_i = 4$ for each of the n units, $n \in \{100, 500, 1000\}$, which also corresponds to the same amount of observations per household available in the EU-SILC data. The part of the predictor, which is common to all types of variables, is of the form

$$1 \cdot X^1_{i,j} - 2 \cdot X^2_{i,j}, \text{ where } X^1_{i,1} = \cdots = X^1_{i,4} \overset{\text{iid}}{\sim} \mathsf{Bernoulli}\,(0.5) \text{ and } X^2_{i,j} \overset{\text{iid}}{\sim} \mathsf{Unif}\,(0,1)\,.$$

Then, we suppose that each unit has its set of observational times $0 < t_{i,1} < t_{i,2} < t_{i,3} < t_{i,4} < 1$ which were generated as an ordered sample from a uniform distribution on an interval (0, 1). We assume the linear parametrization of time. Altogether, the fixed part of the predictor takes the form of $\eta^{\mathsf{F}}_{i,j} = \beta_0 + \beta_1 X^1_{i,j} + \beta_2 X^2_{i,j} + \beta_3 t_{i,j}$. We consider three types of differences assumed among the $G = 2$ or $G = 3$ latent groups:

a) (d = intercept): only the intercept term $\beta_0^{(g)}$ is group-specific, but the slope parameter β_3 is not,

b) (d = slope): only the slope parameters $\beta_3^{(g)}$ is group-specific, but the intercept term β_0 is not,

c) (d = both): both the intercept and the slope terms $\beta_0^{(g)}$, $\beta_3^{(g)}$ are group-specific.

Nevertheless, the model is always estimated under group-specificity of the whole $\boldsymbol{\beta}^{(g)}$ since the implementation for this model could not select only a subset of the fixed effects $\boldsymbol{\beta}$ parameters to be group-specific. This feature is available only for the *GLMM-based* model, see Section 7.1. The values of intercept and slope for each of the nine scenarios were chosen in different ways to obtain clusters distinguishable by the eye (see Figure 5.1).

Another level of scenario settings arise from considering the three types of structures of the random effects:

1. (r = intercept): $\eta^{\mathsf{R}}_{i,j} = b_{0,i}$, random intercept term and fixed slope,
2. (r = slope): $\eta^{\mathsf{R}}_{i,j} = b_{1,i} t_{i,j}$, fixed intercept term and random slope,
3. (r = both): $\eta^{\mathsf{R}}_{i,j} = b_{0,i} + b_{1,i} t_{i,j}$, both intercept and slope are random effects.

We keep the same random effects structure for all outcome types. Therefore, the random effects of i-th unit are multivariate normal of dimension three or six:

$$\boldsymbol{b}_{0,i} = \begin{pmatrix} b^{\mathsf{N}}_{0,i} \\ b^{\mathsf{O}}_{0,i} \\ b^{\mathsf{B}}_{0,i} \end{pmatrix} \sim \mathsf{N}_3\left(\mathbf{0},\, \boldsymbol{\Sigma}_{00}\right), \quad \boldsymbol{b}_{1,i} = \begin{pmatrix} b^{\mathsf{N}}_{1,i} \\ b^{\mathsf{O}}_{1,i} \\ b^{\mathsf{B}}_{1,i} \end{pmatrix} \sim \mathsf{N}_3\left(\mathbf{0},\, \boldsymbol{\Sigma}_{11}\right),$$

$$\boldsymbol{b}_{i} = \begin{pmatrix} \boldsymbol{b}_{0,i} \\ \boldsymbol{b}_{1,i} \end{pmatrix} \sim \mathsf{N}_6\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{01} \\ \boldsymbol{\Sigma}_{10} & \boldsymbol{\Sigma}_{11} \end{pmatrix}\right),$$

for the three scenarios, respectively. Blocks of parameter $\boldsymbol{\Sigma}$ were chosen in the following way:

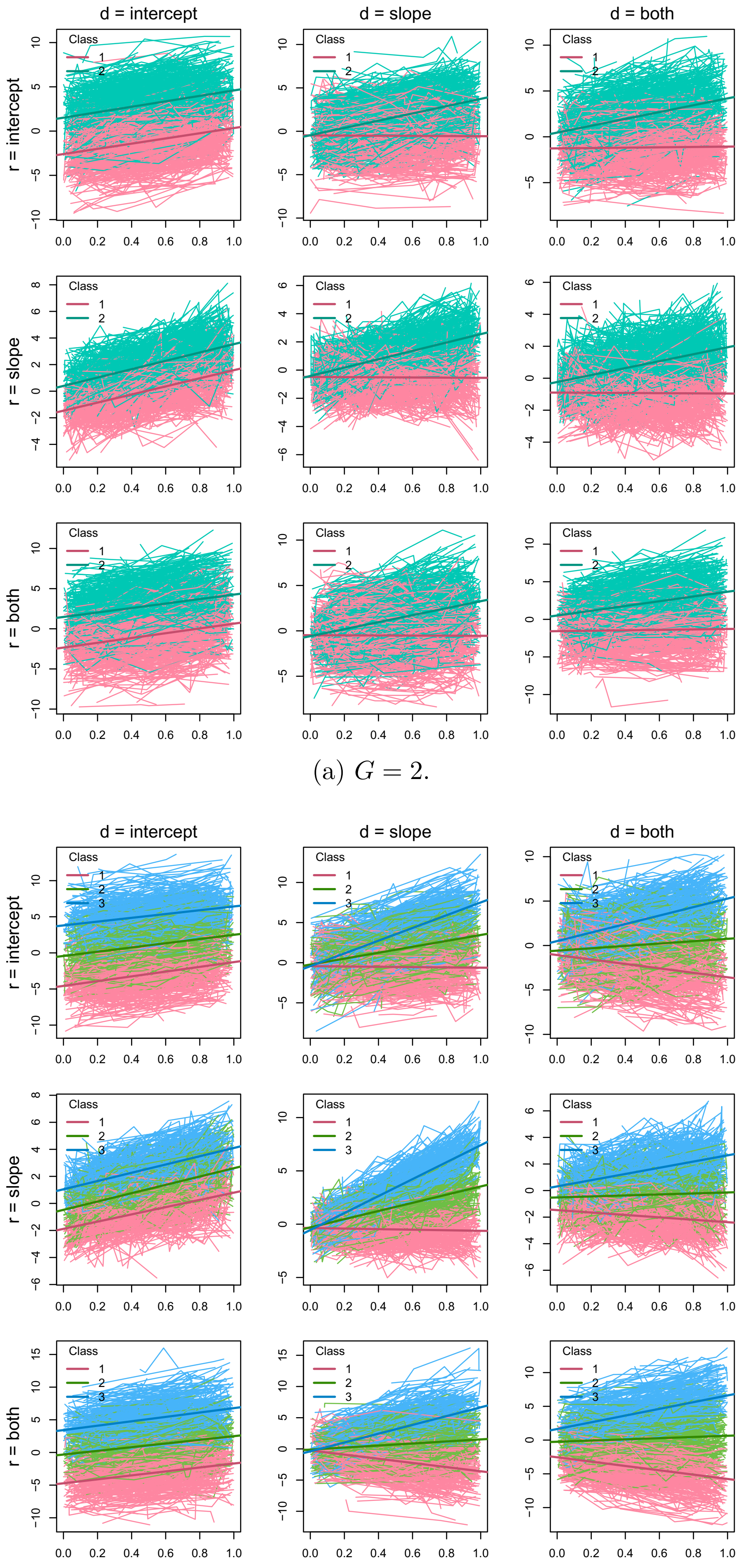
$$\boldsymbol{\Sigma}_{00} = \begin{pmatrix} 2.0 & & \\ & 1.6 & \\ & & 1.2 \end{pmatrix} \begin{pmatrix} 1.0 & 0.6 & 0.6 \\ 0.6 & 1.0 & 0.6 \\ 0.6 & 0.6 & 1.0 \end{pmatrix} \begin{pmatrix} 2.0 & & \\ & 1.6 & \\ & & 1.2 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{11} = \begin{pmatrix} 0.8 & & \\ & 0.6 & \\ & & 0.4 \end{pmatrix} \begin{pmatrix} 1.0 & 0.4 & 0.4 \\ 0.4 & 1.0 & 0.4 \\ 0.4 & 0.4 & 1.0 \end{pmatrix} \begin{pmatrix} 0.8 & & \\ & 0.6 & \\ & & 0.4 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{01} = \boldsymbol{\Sigma}_{10}^{\top} = \begin{pmatrix} 2.0 & & \\ & 1.6 & \\ & & 1.2 \end{pmatrix} \begin{pmatrix} 0.5 & 0.2 & 0.2 \\ 0.2 & 0.5 & 0.2 \\ 0.2 & 0.2 & 0.5 \end{pmatrix} \begin{pmatrix} 0.8 & & \\ & 0.6 & \\ & & 0.4 \end{pmatrix}.$$

These three types of random effects structure are combined with the three types of differences leading to nine different scenarios that are examined for $G = 2, 3$ and different sample sizes n. The group allocation indicator U_i was always generated from a uniform distribution, which results in clusters of comparable sizes. All (latent) numeric outcomes were sampled with unit variance $\tau = 1$. The binary variable was obtained by threshold $\gamma^{\mathsf{B}}_0 = 1$ and the ordinal variable by thresholds $\gamma^{\mathsf{O}}_0 = -1$ and $\gamma^{\mathsf{O}}_1 = 2$. Therefore, if estimated under $\gamma^{\mathsf{B}}_0 = \gamma^{\mathsf{O}}_0 = 0$ the estimates for latent numeric outcomes and γ^{O}_1 should be accordingly shifted.

Each scenario under given G and n was replicated 200-times to explore the properties of the resulting estimators and the classification procedure. For each dataset, the inference is based on an MCMC sample of size $M = 10\,000$. The classification probabilities were calculated for a thinned (1:10) sample to save on the computational time needed to evaluate the multivariate normal integrals (7.4). The simulation study was conducted on a computational cluster consisting of CPU units: Intel(R) Xeon(R) CPU E5-2620 v2, 2.10 GHz, 64 GB RAM. The mean computation time for generating a chain of $M = 10\,000$ sampled values followed by a much more demanding computation of 1 000 classification probabilities for all n units would not take less than an hour even for the lowest values of $n = 100$ and $G = 2$ (around 80 minutes). The most challenging combination of $n = 1\,000$ and $G = 3$ took around 1 200 minutes. The number of calls of `pmvnorm` used for the approximation of posterior distribution of classification probabilities (see Section 7.2) seem to influence the computational time the most; the MCMC sampling itself takes only several seconds to complete (about a minute for the most challenging case $n = 1\,000$ and $G = 3$).



(a) $G = 2$.

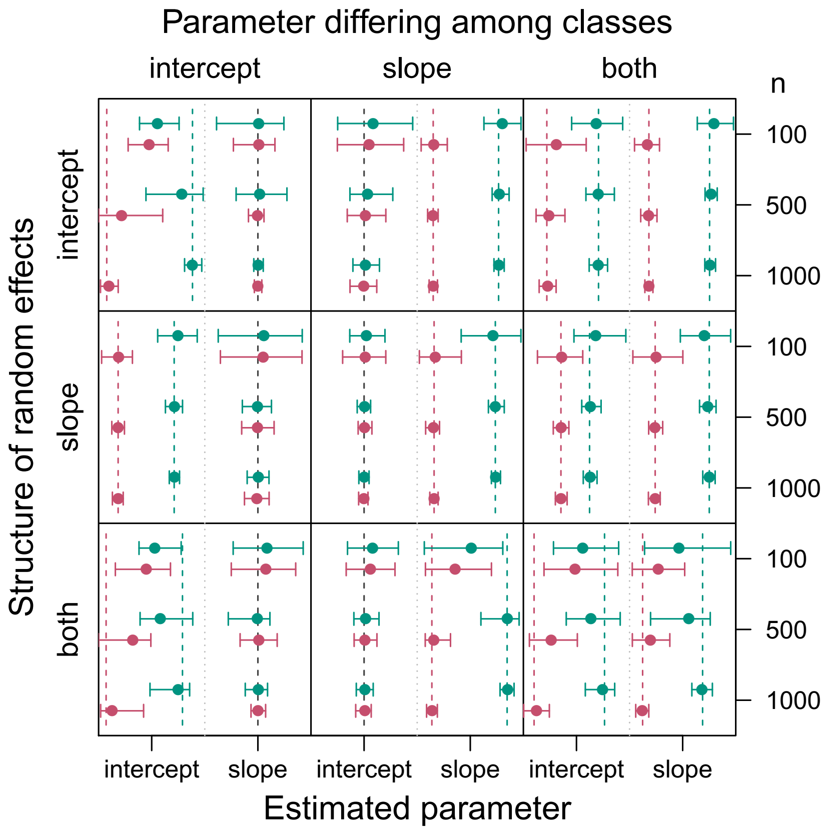
(b) $G = 3$.

Figure 5.1: Simulated samples of a numeric outcome from the *threshold concept* model distinguishing different scenario types (row difference, column structure of random effects).

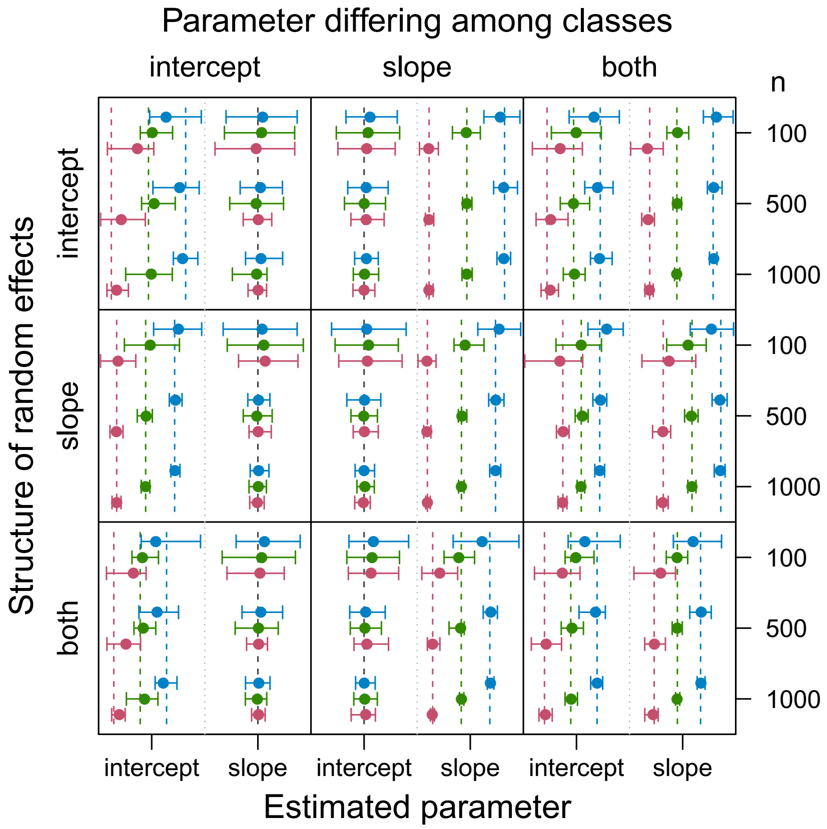
The accuracy of the estimators for the fixed effects

First, Figure 5.2 focuses on the properties of the posterior means of the two main fixed effects (intercept β_0 and slope β_3). The colours distinguish the estimates in different groups ($g = 1, \ldots, G$) and the corresponding true values of the intercept and slope parameters are captured by dashed lines. The dark grey colour depicts the true value shared by all groups. Each segment represents 2.5% and 97.5% quantiles of 200 times replicated estimators and the full circle represents its mean. Figure 5.2 provides estimates of parameters belonging to ordinal outcome only; plots for numeric and binary would depict analogous results.

Figure 5.2 demonstrates that the proposed procedure is capable of providing the estimators with reasonable statistical properties despite the latent modelling and the threshold concept. In most cases, it successfully discovers the difference among groups as intervals of different colours tend not to overlap with each other. There is also an apparent decreasing trend in standard deviation as n increases, suggesting consistency of the estimators. This is disrupted only when the corresponding estimate does not reach the true value. This phenomenon occurs mostly in the estimation of the intercept term when it is considered to be random and different among clusters at the same time. Such behaviour can also be seen for the group-specific slope term when both intercept and slope term are random effects. In these situations, the estimates are shrunk towards the mean of the true values. This might be a result of a combination of the incapability of discrimination between groups for low value of n and the fact that LME usually tends to shrink random effects to zero. In the case of $G = 3$, this effect does not fully vanish even for $n = 1\,000$, see the row *both* and the column *intercept*. However, it seems that sufficiently large number of units n can overcome this issue.



(a) $G = 2$.



(b) $G = 3$.

Figure 5.2: 95% quantile bounds and means for the intercept and slope parameters for the ordinal outcome under different simulation scenarios.

Classification abilities

First, Table 5.1 contains the percentages of the correctly classified units using the HPD interval rule (I2) averaged across the 200 replications. This percentage differs scenario by scenario as the random structures and differences among the groups interact in different ways leading to diverse success rates. For example, the case with a group-specific random slope successfully classifies the vast majority of units for both $G = 2$ and $G = 3$, which is in agreement with the strict separation in the corresponding plots of Figure 5.1. Classification does not work satisfactorily in the problematic cases discussed above in the previous subsection. Since for the low values of n the difference between groups is not estimated to be as strict as it should be, a much larger percentage of units is kept unclassified in such cases. By increasing n, the percentage of unclassified units rapidly decreases and converts mainly into the correctly classified category. Nevertheless, under all scenarios, we managed to keep the misclassification rate very low, always under 10%. The unclassified proportion is also much higher for $G = 3$ as one of the groups (green) is surrounded from both sides, which significantly reduces the ability to distinguish among groups, see Figure 5.1b for illustration.

The classification ability of our approach will also be evaluated by calculating the overall probability that a unit belonging to the cluster g is correctly classified into this cluster. To this end, for each g we calculate the arithmetic mean $\overline{p}_g$ of the MCMC posterior mean estimates $\widehat{U}_{i,g}$ of the allocation probabilities (means of $u_{i,g}(\boldsymbol{\theta}^m), m = 1, \ldots, M$, the evaluation is fully explained in Section 7.2) of belonging to the cluster g across all true cluster members:

$$\overline{p}_g = \frac{1}{|i : U_i = g|} \sum_{i:U_i=g} \widehat{U}_{i,g}. \tag{5.11}$$

Further, to explore the impact of the longitudinally increasing amount of information, we also calculated the classification probabilities “dynamically”. Meaning, that for each unit we pretend a situation that unit i is to be classified on the basis of a set of first $j \in \{1, \ldots, n_i\}$ longitudinal observations that enter the expression (3.6) and consequently also the expression (5.11). Figure 5.3 shows the mean and the quantile bounds of such a dynamically calculated mean probabilities $\overline{p}_2$ based on 200 replications of experiments with $G = 3$ clusters. Group 2 (green) has been chosen for demonstration as it is the middle one that overlaps the other two, which covers the most problematic case (with respect to successful classification).

If a difference among groups lies only in the random intercept term, then there seems to be no improvement with any additional observation. However, in other scenarios, the probability improves with any additional observation from later times as they help to fit the corresponding medium slope value better. This results in rejecting the low and extremely large slope values of other groups, and therefore increasing the probability of classification towards the true middle group. It also improves with the increasing number of units n since the groups are then better distinguished.

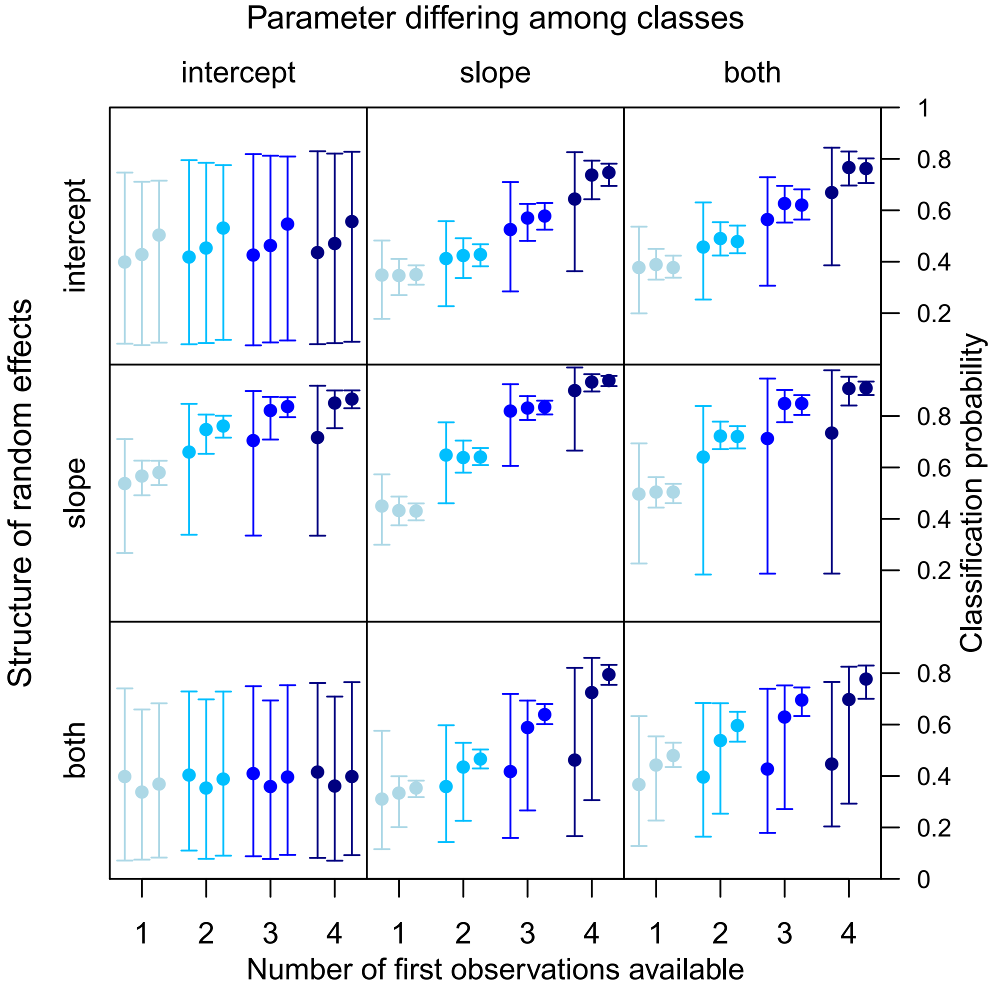


Figure 5.3: Units of group $g = 2$ when $G = 3$. The mean and 2.5% and 97.5% quantiles of mean classification probabilities $\overline{p}_2$ towards the true group calculated dynamically using only first $j \in \{1, 2, 3, 4\}$ observations under several random effects structure and difference among $G = 3$ groups. Three lines of the same colour in one cell correspond to the increasing values of $n \in \{100, 500, 1000\}$.

Table 5.1: Percentages (standard deviation) of correctly classified, unclassified and misclassified units using the HPD interval rule (I2) for several choices of n, G, structure of random effects and group differences in 200 replications.

r[1]	d[2]	n	$G = 2$						$G = 3$					
			Correct [%]		Uncl. [%]		Miscl. [%]		Correct [%]		Uncl. [%]		Miscl. [%]	
.intercept	.intercept	100	27.0	(17.2)	63.2	(25.4)	9.8	(13.7)	23.0	(17.5)	70.2	(21.0)	6.8	(9.4)
		500	62.5	(27.2)	33.0	(27.3)	4.4	(3.8)	44.3	(20.6)	50.8	(22.1)	4.9	(4.4)
		1000	85.1	(6.7)	10.1	(7.1)	4.8	(0.9)	58.6	(16.9)	35.5	(17.2)	6.0	(3.1)
.intercept	slope	100	76.8	(5.4)	20.3	(5.5)	2.9	(1.9)	56.0	(8.5)	40.4	(8.9)	3.6	(2.4)
		500	86.1	(1.8)	8.9	(1.8)	5.0	(1.0)	74.6	(2.0)	19.0	(2.1)	6.4	(1.2)
		1000	87.5	(1.1)	6.7	(0.9)	5.9	(0.7)	78.2	(1.5)	13.8	(1.5)	8.0	(0.8)
.intercept	both	100	86.5	(4.4)	12.0	(4.4)	1.5	(1.1)	58.0	(9.4)	38.5	(10.2)	3.4	(2.2)
		500	92.9	(1.4)	4.5	(1.1)	2.6	(0.7)	76.9	(2.5)	16.5	(2.5)	6.7	(1.1)
		1000	93.8	(0.8)	3.3	(0.6)	2.9	(0.5)	79.4	(1.6)	12.8	(1.6)	7.8	(0.8)
slope	.intercept	100	96.2	(2.6)	3.4	(2.5)	0.4	(0.6)	61.2	(15.5)	36.4	(15.7)	2.3	(1.8)
		500	97.9	(0.5)	1.5	(0.5)	0.6	(0.4)	87.6	(2.2)	9.2	(2.2)	3.2	(0.7)
		1000	98.3	(0.4)	0.9	(0.3)	0.8	(0.3)	90.2	(1.2)	6.2	(1.1)	3.6	(0.5)
slope	slope	100	80.1	(20.4)	16.3	(19.0)	3.6	(8.7)	85.7	(13.5)	13.3	(13.6)	1.0	(1.2)
		500	92.8	(1.5)	4.6	(1.4)	2.6	(0.7)	94.9	(1.2)	3.6	(1.0)	1.5	(0.5)
		1000	93.9	(0.9)	3.3	(0.7)	2.8	(0.5)	95.5	(0.7)	2.6	(0.5)	1.9	(0.4)
slope	both	100	85.3	(18.0)	13.8	(18.0)	0.9	(0.9)	62.2	(23.5)	35.8	(23.4)	2.0	(2.7)
		500	96.2	(1.0)	2.6	(0.9)	1.3	(0.6)	92.4	(1.7)	5.5	(1.5)	2.1	(0.8)
		1000	96.7	(0.6)	1.8	(0.4)	1.5	(0.4)	93.3	(0.9)	4.1	(0.9)	2.5	(0.5)
both	.intercept	100	18.8	(13.7)	76.0	(16.6)	5.2	(7.2)	18.7	(15.2)	78.1	(16.7)	3.2	(4.1)
		500	35.4	(25.2)	58.7	(27.2)	6.0	(8.5)	30.6	(18.5)	65.1	(20.5)	4.3	(3.9)
		1000	70.5	(22.4)	24.3	(23.4)	5.2	(1.9)	46.4	(12.1)	48.2	(13.9)	5.4	(2.4)
both	slope	100	16.2	(13.2)	79.2	(16.8)	4.5	(6.0)	23.4	(22.3)	74.9	(23.6)	1.6	(2.3)
		500	69.7	(18.1)	24.7	(19.4)	5.6	(2.0)	69.8	(13.4)	25.2	(14.4)	5.0	(1.4)
		1000	80.5	(3.0)	12.0	(3.3)	7.4	(1.2)	81.1	(2.2)	11.9	(2.1)	7.0	(0.8)
both	both	100	16.7	(14.5)	80.3	(17.3)	3.0	(5.5)	19.4	(19.8)	79.7	(20.7)	0.9	(1.4)
		500	43.6	(30.5)	53.3	(32.3)	3.0	(2.8)	66.3	(19.6)	29.1	(21.1)	4.5	(1.9)
		1000	80.3	(10.5)	13.5	(11.1)	6.2	(1.2)	80.9	(3.3)	12.1	(3.5)	7.0	(1.0)

[1] Structure of random effects.

[2] Difference among groups.

6. MCMC estimation of the GLMM-based model

This chapter is dedicated solely to the *GLMM-based* model (Sections 2.4.2, 3.3.2, 4.4.2) and derivation of all necessary quantities to construct an MCMC algorithm. Similarly as for the *threshold concept* model we aim for the Gibbs sampler, although this time we are unable to classify some full-conditionals into the well-known distributional families or to directly sample from the full-conditional distribution. Hence, we start with the achievable full-conditionals and list down the parameters that require the help of *Metropolis proposal* step. Such a combination of the two sampling techniques is known as *Metropolis within Gibbs sampler*.

Then, we focus on the proposal distribution itself. The algorithm benefits from generating proposals from distribution of similar characteristics, therefore, the goal is to find a reasonable approximation of the target distribution. We describe, how *Taylor expansion* and *Newton–Raphson method* (Algorithm 6) are used in this context.

This model is more advanced than the previous one, mainly in the use of *sparse finite mixture* methodology (Malsiner-Walli et al., 2016; Frühwirth-Schnatter and Malsiner-Walli, 2019) to estimate the number of clusters. Simply put, the number of non-empty clusters decreases from its maximal amount G_{max} during the sampling until a reasonable count is achieved. Group-specific parameters are sampled for all G_{max} clusters nonetheless. For that reason, the sampling algorithm has to be adjusted and accompanied with a post-sampling procedure to process the sampled data and deal with potential *label switching problems*.

Finally, a simulation study is performed to test the ability to estimate the model parameters as well as the true number of underlying clusters.

6.1 Full-conditional distributions

The implementation for this model is more advanced in yet another aspect. The data usually lack some outcome values, therefore, one datarow may be completely ignored even if just a single value is missing since the implementation of the *threshold concept* model works under the complete-case analysis. Here, we are able to bypass this problem with the use of BDA similarly as with any other unobserved model quantity, see Section 4.1.

Let us divide $\mathbb{Y} = \left\{\mathbb{Y}^{\mathsf{obs}}, \mathbb{Y}^{\mathsf{mis}}\right\}$, where $\mathbb{Y}^{\mathsf{obs}}$ is the observed part of the outcomes, while $\mathbb{Y}^{\mathsf{mis}}$ are the unobserved (missing) outcome values. At some places, this division will be crucial, however, when not, the traditional symbol $\mathbb{Y}$ is used, e.g. full-conditional distributions of other parameters are derived given the whole $\mathbb{Y}$. By BDA, the set of latent quantities is then extended to $\mathcal{L} = \{\mathbb{Y}^{\mathsf{mis}}, U_i, \boldsymbol{b}_i;\ i = 1, \dots, n\}$. For row (i, j), at least some outcome value has to be observed to contribute to the model. Otherwise, only the predictive distribution of $\boldsymbol{Y}_{i,j}$ given covariates $\mathcal{C}_{i,j}$ is explored.

For this model, the set of all randomized elements (including the latent variables) of the model $\boldsymbol{\Psi}$ consists of $\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{c}, \boldsymbol{\Sigma}, \mathbb{Y}^{\mathsf{mis}}, \boldsymbol{U}, \boldsymbol{b}, \mathbb{Q}$ and e_0. Again, to derive full-conditional distributions for all parameters $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, we have to view the right hand side of (4.20) as a function of parameter $\boldsymbol{\psi}$ which can be decomposed into the following products:

$$
\begin{aligned}
&p\left(\boldsymbol{w},\, \boldsymbol{\beta},\, \boldsymbol{\tau},\, \boldsymbol{c},\, \boldsymbol{\Sigma},\, \mathbb{Y}^{\mathsf{mis}},\, \boldsymbol{U},\, \boldsymbol{b},\, \mathbb{Q},\, e_0 \,\middle|\, \mathbb{Y}^{\mathsf{obs}};\, \mathcal{C}\right) \propto \\
&\propto \prod_{i=1}^{n} \left[\prod_{r \in \mathcal{R}} \prod_{j=1}^{n_i} p_{\mathsf{t}(r)} \left(Y_{i,j}^{r} \,\middle|\, \boldsymbol{b}_i^{r}, \boldsymbol{\beta}_r^{(U_i)}, \tau_r^{(U_i)}, \boldsymbol{c}_r^{(U_i)}; \mathcal{C}_{i,j} \right) \cdot p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\Sigma}^{(U_i)}\right) \cdot p(U_i|\boldsymbol{w}) \right] \cdot \\
&\cdot p(\boldsymbol{w}|e_0)p(e_0|a_e, b_e)p(\boldsymbol{c}|\boldsymbol{\alpha})p(\boldsymbol{\beta}\,|\,\boldsymbol{\tau}, \boldsymbol{\beta}_0, \mathbb{D})p(\boldsymbol{\tau}|a_\tau, b_\tau)p(\boldsymbol{\Sigma}|\mathbb{Q}, \nu_0)p(\mathbb{Q}|\mathbb{D}_{\mathbb{Q}}, \nu_1), \qquad (6.1)
\end{aligned}
$$

where $\mathcal{H}_0$ denotes all fixed hyperparameters of prior distributions. Similarly as before, the latent quantities remain present compared to (3.5) instead of being integrated, which considerably simplifies the evaluation. Combining the arrows in Figure 4.4 coming in and out of the node of interest $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, we can determine the factors needed for the full-conditional distribution of $\boldsymbol{\psi}$.

The following is a list of parameters with the full-conditional distribution of the same shape as in Section 5.1:

- the cluster allocation probabilities $\boldsymbol{w}$, equation (5.2),
- precision parameters $\tau_r^{(g)}$ for numeric outcomes $r \in \mathcal{R}^{\mathsf{Num}}$, equation (5.6),
- fixed effects $\boldsymbol{\beta}_r^{(g)}$ corresponding to numeric outcomes $r \in \mathcal{R}^{\mathsf{Num}}$, equation (5.7),
- prior scale matrix $\mathbb{Q}$ for matrices $\boldsymbol{\Sigma}$, equation (5.8),
- covariance matrices $\boldsymbol{\Sigma}^{(g)}$ for random effects[*], equation (5.9).

Missing outcome values

Missing outcome values $\mathbb{Y}^{\mathsf{mis}}$ have the most trivial full-conditional distribution among all $\boldsymbol{\psi} \in \boldsymbol{\Psi}$. One simply has to sample an unobserved $Y_{i,j}^{r}$ according to the corresponding GLMM using the group-specific parameters of cluster $g = U_i$. This is possible since the random effects $\boldsymbol{b}_i^r$ are known. In general, the following notation could be used

$$
Y_{i,j}^{r} \,\middle|\, U_i = g, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r^{(g)}, \tau_r^{(g)}, \boldsymbol{c}_r^{(g)}; \mathcal{C}_{i,j} \;\sim\; p_{\mathsf{t}(r)} \left(Y_{i,j}^{r} \,\middle|\, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r^{(g)}, \tau_r^{(g)}, \boldsymbol{c}_r^{(g)}; \mathcal{C}_{i,j} \right), \qquad (6.2)
$$

where the usual symbol for pdf $p_{\mathsf{t}(r)}$ now stands for the assumed model for outcome of type $\mathsf{t}(r)$, which is given by either (2.1), (2.7), (2.9), (2.12) or (2.14).

Full-conditional clustering probabilities

We can derive the full-conditional clustering probabilities easily the same way as in Section 5.1.2. Simply take all factors where any group-specific parameter is

[*] Even though the random effects $\boldsymbol{b}_i$ now contain effects for many more types of outcomes, compare Sections 2.4.1 and 2.4.2, they can still be vectorized into $\boldsymbol{b}_i$ and the derivation proceeds in the same manner as in Section 5.1.8.

chosen according to U_i and its prior distribution:

$$\begin{aligned} &p\left(U_i \,\middle|\, \mathbb{Y}, \boldsymbol{\Psi}_{-U_i}, \mathcal{H}_0; \mathcal{C}\right) \propto \\ &\qquad \propto \prod_{r\in\mathcal{R}} \prod_{j=1}^{n_i} p_{\mathsf{t}(r)}\left(Y^r_{i,j} \,\middle|\, \boldsymbol{b}^r_i, \boldsymbol{\beta}_r^{(U_i)}, \tau_r^{(U_i)}, \boldsymbol{c}_r^{(U_i)}; \mathcal{C}_{i,j}\right) \cdot p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\Sigma}^{(U_i)}\right) \cdot p\left(U_i | \boldsymbol{w}\right). \end{aligned}$$

The full-conditional probability of belonging to cluster $g = 1, \dots, G$ could be expressed in detail using (2.24) as

$$\begin{aligned} \mathsf{P}\left[U_i = g \,|\, \mathbb{Y}_i, \boldsymbol{b}_i, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{c}, \boldsymbol{\Sigma}; \mathcal{C}_i\right] &\propto w_g \cdot \left|\boldsymbol{\Sigma}^{-(g)}\right|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-(g)} \boldsymbol{b}_i\right\} \cdot \\ &\cdot \prod_{r\in\mathcal{R}^{\mathsf{Num}}} \left(\tau_r^{(g)}\right)^{\frac{n_i}{2}} \exp\left\{-\frac{\tau_r^{(g)}}{2} \sum_{j=1}^{n_i} \left(Y^r_{i,j} - \eta^{r,(g)}_{i,j}\right)^2\right\} \cdot \\ &\cdot \prod_{r\in\mathcal{R}^{\mathsf{Poi}}} \exp\left\{\sum_{j=1}^{n_i} \left(Y^r_{i,j}\eta^{r,(g)}_{i,j} - \exp\left\{\eta^{r,(g)}_{i,j}\right\}\right)\right\} \cdot \\ &\cdot \prod_{r\in\mathcal{R}^{\mathsf{Bin}}} \prod_{j=1}^{n_i} \left[\mathsf{logit}^{-1}\left(\eta^{r,(g)}_{i,j}\right)\right]^{Y^r_{i,j}} \left[1 - \mathsf{logit}^{-1}\left(\eta^{r,(g)}_{i,j}\right)\right]^{1-Y^r_{i,j}} \cdot \\ &\cdot \prod_{r\in\mathcal{R}^{\mathsf{Ord}}} \prod_{j=1}^{n_i} \left[\mathsf{logit}^{-1}\left(\eta^{r,(g)}_{i,j} - c^{(g)}_{r,Y^r_{i,j}-1}\right) - \mathsf{logit}^{-1}\left(\eta^{r,(g)}_{i,j} - c^{(g)}_{r,Y^r_{i,j}}\right)\right] \cdot \\ &\cdot \prod_{r\in\mathcal{R}^{\mathsf{Cat}}} \prod_{j=1}^{n_i} \frac{\exp\left\{\eta^{r,(g)}_{i,j,Y^r_{i,j}}\right\}}{1 + \sum_{k=1}^{K^r-1} \exp\{\eta^{r,(g)}_{i,j,k}\}}, \end{aligned} \tag{6.3}$$

where $\eta^{r,(g)}_{i,j} = \left(\boldsymbol{x}^r_{i,j}\right)^\top \boldsymbol{\beta}^{(g)}_r + \left(\boldsymbol{z}^r_{i,j}\right)^\top \boldsymbol{b}^r_i$ is the linear predictor of j-th observation of outcome $r \in \mathcal{R} \setminus \mathcal{R}^{\mathsf{Cat}}$ of unit i when belonging to the group g. We denote the predictor for categorical outcomes in similar fashion but with an additional subscript k denoting the level to which the predictor corresponds.

From the computational point of view, it is more convenient to calculate the logarithm of the right hand side up to any multiplicative constant. In this case, we arrive to

$$\begin{aligned} \log \mathsf{P}\left[U_i = g \,|\, \mathbb{Y}_i, \boldsymbol{b}_i, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{c}, \boldsymbol{\Sigma}; \mathcal{C}_i\right] &= \text{ const. } + \log w_g + \frac{1}{2}\log\left|\boldsymbol{\Sigma}^{-(g)}\right| - \\ &- \frac{1}{2}\boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-(g)} \boldsymbol{b}_i + \sum_{r\in\mathcal{R}^{\mathsf{Num}}} \frac{n_i}{2}\log\left(\tau_r^{(g)}\right) - \frac{\tau_r^{(g)}}{2}\sum_{j=1}^{n_i}\left(Y^r_{i,j} - \eta^{r,(g)}_{i,j}\right)^2 + \\ &+ \sum_{r\in\mathcal{R}^{\mathsf{Poi}}} \sum_{j=1}^{n_i} \left(Y^r_{i,j}\eta^{r,(g)}_{i,j} - \exp\left\{\eta^{r,(g)}_{i,j}\right\}\right) + \\ &+ \sum_{r\in\mathcal{R}^{\mathsf{Bin}}} \sum_{j=1}^{n_i} \left[Y^r_{i,j}\eta^{r,(g)}_{i,j} - \log\left(1 + \exp\left\{\eta^{r,(g)}_{i,j}\right\}\right)\right] + \\ &+ \sum_{r\in\mathcal{R}^{\mathsf{Ord}}} \sum_{j=1}^{n_i} \log\left[\mathsf{logit}^{-1}\left(\eta^{r,(g)}_{i,j} - c^{(g)}_{r,Y^r_{i,j}-1}\right) - \mathsf{logit}^{-1}\left(\eta^{r,(g)}_{i,j} - c^{(g)}_{r,Y^r_{i,j}}\right)\right] + \\ &+ \sum_{r\in\mathcal{R}^{\mathsf{Cat}}} \sum_{j=1}^{n_i} \left[\eta^{r,(g)}_{i,j,Y^r_{i,j}} - \log\left(1 + \sum_{k=1}^{K^r-1} \exp\left\{\eta^{r,(g)}_{i,j,k}\right\}\right)\right]. \end{aligned} \tag{6.4}$$

Having computed the right hand side we proceed as described in Section 5.1.2.

Hyperparameter e_0

Mixture weights $\boldsymbol{w}$ are apriori given by the Dirichlet prior (4.11) controlled by hyperparameter e_0. Its value is regulated by Gamma prior (4.12) with carefully chosen shape a_e and rate b_e. The full-conditional pdf is then proportional to

$$p\left(e_0|\boldsymbol{w}, a_e, b_e\right) \propto p(\boldsymbol{w}|e_0)p(e_0|a_e, b_e) =\propto \frac{\Gamma(Ge_0)}{\Gamma^G(e_0)} \prod_{g=1}^{G} w_g^{e_0-1} \cdot e_0^{a_e-1} \exp\{-b_e e_0\}. \tag{6.5}$$

Here we do not immediately recognize the full-conditional distribution, hence, a Metropolis proposal needs to be employed.

We follow the steps of Frühwirth-Schnatter and Malsiner-Walli (2019) and sample e_0 from partly marginalized full-conditional distribution instead. In particular, we integrate the parameter $\boldsymbol{w}$ out of the conditioning, hence, we sample e_0 from $p(e_0|\,\boldsymbol{U}) \propto p(e_0)p(\boldsymbol{U}|\,e_0)$ instead of $p(e_0|\,\boldsymbol{w})$.

Parameter $\boldsymbol{w}$ has to be integrated out of the combination of (3.1) and (4.11)

$$p(\boldsymbol{U},\, \boldsymbol{w}\,|\,e_0\,) = \frac{\Gamma(Ge_0)}{\Gamma^G(e_0)} \cdot \prod_{g=1}^{G} w_g^{n^{(g)}(\boldsymbol{U})+e_0-1}.$$

The so called *stick-breaking representation*

$$\begin{aligned}
w_1 &= v_1, & v_1 &= w_1,\\
w_2 &= v_2(1-v_1), & v_2 &= w_2/(1-w_1),\\
w_3 &= v_3(1-v_2)(1-v_1), & v_3 &= w_3/(1-w_1-w_2),\\
\vdots &= \vdots & \vdots &= \vdots\\
w_g &= v_g \prod_{j=1}^{g-1}(1-v_j), & v_g &= w_g \Big/ \left(1-\sum_{j=1}^{g-1} w_g\right)
\end{aligned} \tag{6.6}$$

and $v_G = 1$ transforms the integration with respect to (wrt) $\boldsymbol{w}$ into

$$\int \prod_{g=1}^{G} w_g^{n^{(g)}(\boldsymbol{U})+e_0-1}\, \mathrm{d}\boldsymbol{w} = \int \prod_{g=1}^{G-1} v_g^{n^{(g)}(\boldsymbol{U})+e_0-1} (1-v_g)^{(G-g)e_0 + \sum\limits_{g'=g+1}^{G} n^{(g')}(\boldsymbol{U})-1}\, \mathrm{d}\boldsymbol{v},$$

which can be decomposed into $G-1$ integrals wrt independent $v_g \in (0,1)$. We immediately recognize the shape of Beta distributions

$$\mathsf{Beta}\left(n^{(g)}(\boldsymbol{U}) + e_0, (G-g)e_0 + \sum_{g'=g+1}^{G} n^{(g')}(\boldsymbol{U})\right),$$

hence, the integral corresponds to product of beta functions. After evaluation and simplification we arrive to

$$p(\boldsymbol{U}\,|\,e_0\,) = \frac{\Gamma(Ge_0)}{\Gamma(Ge_0+n)} \prod_{g:n^{(g)}(\boldsymbol{U})>0} \frac{\Gamma(n^{(g)}(\boldsymbol{U})+e_0)}{\Gamma(e_0)}. \tag{6.7}$$

Combining (6.7) with pdf (4.12) of prior for e_0 we obtain

$$p(e_0|\,\boldsymbol{U}) \propto e_0^{a_e-1}\exp\{-b_e e_0\} \cdot \frac{\Gamma(Ge_0)}{\Gamma(Ge_0+n)} \prod_{g:n^{(g)}(\boldsymbol{U})>0} \frac{\Gamma(n^{(g)}(\boldsymbol{U})+e_0)}{\Gamma(e_0)}. \tag{6.8}$$

6.2 Metropolis proposal steps

Within the MCMC estimation procedure, we also need to sample from full-conditional (or partly marginalized) distributions of parameters which do not fall into well-known distributional families. This complicates the sampling.

In the following we assume that we work with a parameter $\boldsymbol{\omega} \in \mathbb{R}^{\kappa}$ from which we want to sample with respect to a distribution given by a pdf proportional to a twice differentiable function $p(\boldsymbol{\omega}) > 0, \forall \omega \in \mathbb{R}^{\kappa}$. This differentiability property is also transferred to the corresponding log-pdf $\ell(\boldsymbol{\omega}) = \log p(\boldsymbol{\omega})$ that can be arbitrarily shifted by a constant.

Given a previous value $\boldsymbol{\omega}^m$ we want to find a suitable proposal $\boldsymbol{\omega}^{m+1}$ for the next value of the parameter $\boldsymbol{\omega}$. We adopt a random walk approach with independent steps sampled from a centred multivariate normal distribution with variance matrix $\boldsymbol{\Omega}$, i.e. $\boldsymbol{\omega}^{m+1} \sim \mathsf{N}_{\kappa}\left(\boldsymbol{\omega}^m, \boldsymbol{\Omega}\right)$. The proposal $\boldsymbol{\omega}^{m+1}$ is then accepted with probability

$$\begin{aligned}\alpha\left(\boldsymbol{\omega}^{m+1}, \boldsymbol{\omega}^m\right) &= \min\left\{1, \frac{p\left(\boldsymbol{\omega}^{m+1}\right)}{p\left(\boldsymbol{\omega}^m\right)}\right\} \\ &= \begin{cases} \exp\left\{\ell\left(\boldsymbol{\omega}^{m+1}\right) - \ell\left(\boldsymbol{\omega}^m\right)\right\}, & \text{if } \ell\left(\boldsymbol{\omega}^{m+1}\right) < \ell\left(\boldsymbol{\omega}^m\right), \\ 1, & \text{if } \ell\left(\boldsymbol{\omega}^{m+1}\right) \geq \ell\left(\boldsymbol{\omega}^m\right). \end{cases}\end{aligned}$$

The suitable choice of the variance matrix $\boldsymbol{\Omega}$ is crucial as a poor choice results in an inappropriate exploration of the posterior.

Using the *Taylor expansion* at $\widehat{\boldsymbol{\omega}}$ maximising the (log-)pdf, thus satisfying $\frac{\partial}{\partial \omega}\ell\left(\widehat{\omega}\right) = \mathbf{0}$, we obtain the following approximation

$$\ell\left(\boldsymbol{\omega}\right) \;\approx\; \text{const.} - \frac{1}{2}\left(\boldsymbol{\omega} - \widehat{\boldsymbol{\omega}}\right)^{\top}\left[-\frac{\partial^2 \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^{\top}}\bigg|_{\omega=\widehat{\omega}}\right]\left(\boldsymbol{\omega} - \widehat{\boldsymbol{\omega}}\right).$$

Hence, we want to sample from the pdf which locally (around $\widehat{\boldsymbol{\omega}}$) resembles the pdf of $\mathsf{N}_{\kappa}\left(\widehat{\boldsymbol{\omega}}, \left[-\frac{\partial^2 \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^{\top}}\bigg|_{\omega=\widehat{\omega}}\right]^{-1}\right)$. Hence, we use the variance matrix $\boldsymbol{\Omega} = c_{\omega} \cdot \left[-\frac{\partial^2 \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^{\top}}\bigg|_{\omega=\widehat{\omega}}\right]^{-1}$ for the multivariate normal distribution to sample the increment when proposing a new value of $\boldsymbol{\omega}$. The multiplicative constant c_{ω} (close to 1) is used to shrink or stretch the size of the increment steps.

This matrix does not have to be updated in every iteration m. Especially, once the limiting distribution of the chain is reached, $\boldsymbol{\Omega}$ should be more or less the same and hence should be updated rarely to save computational time. We also propose several transitions between $\boldsymbol{\omega}^{m+1}$ and $\boldsymbol{\omega}^m$ to speed up convergence to the limiting distribution and to make better use of the costly computation of $\boldsymbol{\Omega}$.

To find $\widehat{\boldsymbol{\omega}}$ maximising $\ell(\boldsymbol{\omega})$, we employ the *Newton–Raphson method.* Starting from some initial value $\boldsymbol{\omega}_0$, e.g. the maximum from the previous step, we iteratively solve

$$\left[-\frac{\partial^2 \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^{\top}}\bigg|_{\omega=\omega_k}\right]\boldsymbol{s} = \frac{\partial \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\bigg|_{\omega=\omega_k}.$$

to find the direction in which to move from current position $\boldsymbol{\omega}_k$, see Algorithm 6 in Section 7.4 for more details. Hence, evaluation of the gradient and Hess matrix has to be feasible. This algorithm primarily yields the basis for the precision matrix $\boldsymbol{\Omega}^{-1}$ of the incremental distribution.

In the following subsections we explore in detail the peculiarities of individual parameters that require a Metropolis proposal approach for sampling from the full-conditional distribution. These include: the log-precision $e_0^\star := \log e_0$ (to sample $e_0 > 0$), the fixed effects $\boldsymbol{\beta}_r^{(g)}$ of non-numeric outcomes $r \in \mathcal{R} \setminus \mathcal{R}^{\mathsf{Num}}$, the random effects $\boldsymbol{b}_i$ specific to each unit i and the transformed ordered intercepts $\boldsymbol{a}_r^{(g)} = a\left(\boldsymbol{c}_r^{(g)}\right)$ (to sample the ordered intercepts $\boldsymbol{c}_r^{(g)}$ and the corresponding probabilities $\boldsymbol{\pi}_r^{(g)}$). We derive in detail the corresponding ℓ functions together with gradient and negative Hess matrix, which are required by Algorithm 6 to find a suitable proposal matrix $\boldsymbol{\Omega}$ for each of these parameters.

6.2.1 Hyperparameter e_0

Since e_0 is positive, we will perform the proposal on a log-scale by defining a new parameter $e_0^\star = \log e_0 \in \mathbb{R}$, which by the transformation theorem yields

$$\begin{aligned} p(e_0^\star \,|\, \boldsymbol{U}\,) \propto \exp\{e_0^\star a_e - b_e \exp\{e_0^\star\}\}\cdot& \\ \cdot \frac{\Gamma(G\exp\{e_0^\star\})}{\Gamma(G\exp\{e_0^\star\}+n)} \prod_{g:n^{(g)}(\boldsymbol{U})>0} \frac{\Gamma(n^{(g)}(\boldsymbol{U})+\exp\{e_0^\star\})}{\Gamma(\exp\{e_0^\star\})}.& \qquad (6.9) \end{aligned}$$

Transforming (6.9) into log-scale yields

$$\begin{aligned} \ell(e_0^\star \,|\, \boldsymbol{U}\,) = {} & \text{const.} + e_0^\star a_e - b_e \exp\{e_0^\star\} + \log\Gamma(G\exp\{e_0^\star\}) \\ & - \log\Gamma(G\exp\{e_0^\star\}+n) + \sum_{\{g:n^{(g)}(\boldsymbol{U})>0\}} \log\Gamma(n^{(g)}(\boldsymbol{U})+\exp\{e_0^\star\}) \\ & - G_+\log\Gamma(\exp\{e_0^\star\}). \qquad (6.10) \end{aligned}$$

The first and second derivative of (6.10) can be obtained with the use of the derivatives of the log-gamma function $\log\Gamma$, namely the digamma function ψ and the trigamma function ψ_1, both implemented in base R. They take the following form:

$$\begin{aligned} [\star] &= -b_e + G\psi(G\exp\{e_0^\star\}) - G\psi(G\exp\{e_0^\star\}+n) \\ &\quad + \sum_{\{g:n^{(g)}(\boldsymbol{U})>0\}} \psi(n^{(g)}(\boldsymbol{U})+\exp\{e_0^\star\}) - G_+\psi(\exp\{e_0^\star\}), \\ [*] &= G^2\psi_1(G\exp\{e_0^\star\}) - G^2\psi_1(G\exp\{e_0^\star\}+n) \\ &\quad + \sum_{\{g:n^{(g)}(\boldsymbol{U})>0\}} \psi_1(n^{(g)}(\boldsymbol{U})+\exp\{e_0^\star\}) - G_+\psi_1(\exp\{e_0^\star\}), \\ \frac{\partial\ell(e_0^\star|\,\boldsymbol{U})}{\partial e_0^\star} &= a_e + \exp\{e_0^\star\}\cdot[\star], \\ \frac{\partial^2\ell(e_0^\star|\,\boldsymbol{U})}{\partial\left(e_0^\star\right)^2} &= \exp\{e_0^\star\}\cdot\left([\star] + \exp\{e_0^\star\}[*]\right). \end{aligned}$$

The new $e_0^\star$ is proposed using this combination of a Newton–Raphson step and a random walk and if accepted, we transform it back to obtain the new $e_0 = \exp\{e_0^\star\}$.

6.2.2 Fixed effects $\boldsymbol{\beta}_r$ for non-numeric outcomes

Table 6.1 contains an overview of the contributions of a single outcome observation to the log-likelihood depending on the type of the outcome. Moreover, derivatives with respect to the predictor η (or $\boldsymbol{\eta}$) can be further used for determining the derivatives with respect to fixed and random effects. In this section, which is devoted to the fixed effects, we will use that

$$\frac{\partial \eta}{\partial \boldsymbol{\beta}} = \frac{\partial\left(\boldsymbol{x}^\top \boldsymbol{\beta} + \eta^{\mathsf{R}}\right)}{\partial \boldsymbol{\beta}} = \boldsymbol{x},$$

where η^{R} denotes the random-effects part of the linear predictor.

In the following, we present the log-posteriors and their derivatives for the full-conditional distribution of the fixed effects $\boldsymbol{\beta}_r^{(g)}$ within the group g for a count, binary, ordinal and general categorical outcome. We kindly remind the notation $\eta_{i,j}^{r,(g)}$ for the predictor formed by $\boldsymbol{\beta}_r^{(g)}$. We start with a count outcome, $r \in \mathcal{R}^{\mathsf{Poi}}$:

$$\begin{aligned}
\ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right) &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \left[Y_{i,j}^r \eta_{i,j}^{r,(g)} - \exp\left\{\eta_{i,j}^{r,(g)}\right\}\right] \\
&\quad - \frac{1}{2}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right), \\
\frac{\partial \ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right)}{\partial \boldsymbol{\beta}_r^{(g)}} &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \left[Y_{i,j}^r - \exp\left\{\eta_{i,j}^{r,(g)}\right\}\right] \boldsymbol{x}_{i,j}^r \\
&\quad - \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right), \\
-\frac{\partial^2 \ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right)}{\partial \boldsymbol{\beta}_r^{(g)} \partial\left(\boldsymbol{\beta}_r^{(g)}\right)^\top} &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \exp\left\{\eta_{i,j}^{r,(g)}\right\} \boldsymbol{x}_{i,j}^r \left(\boldsymbol{x}_{i,j}^r\right)^\top + \mathbb{D}_r^{-1}.
\end{aligned}$$

We continue with a binary outcome, $r \in \mathcal{R}^{\mathsf{Bin}}$:

$$\begin{aligned}
\ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right) &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \left[Y_{i,j}^r \eta_{i,j}^{r,(g)} - \log\left(1 + \exp\left\{\eta_{i,j}^{r,(g)}\right\}\right)\right] \\
&\quad - \frac{1}{2}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right), \\
\frac{\partial \ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right)}{\partial \boldsymbol{\beta}_r^{(g)}} &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \left[Y_{i,j}^r - \mathsf{logit}^{-1}\left(\eta_{i,j}^{r,(g)}\right)\right] \boldsymbol{x}_{i,j}^r \\
&\quad - \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right), \\
-\frac{\partial^2 \ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right)}{\partial \boldsymbol{\beta}_r^{(g)} \partial\left(\boldsymbol{\beta}_r^{(g)}\right)^\top} &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \left[\mathsf{logit}^{-1}\left(\eta_{i,j}^{r,(g)}\right) \cdot \right. \\
&\quad \left. \left(1 - \mathsf{logit}^{-1}\left(\eta_{i,j}^{r,(g)}\right)\right)\right] \boldsymbol{x}_{i,j}^r \left(\boldsymbol{x}_{i,j}^r\right)^\top + \mathbb{D}_r^{-1}.
\end{aligned}$$

Next, the log-posterior and its derivatives of the full-conditional distribution of the fixed effects $\boldsymbol{\beta}_r^{(g)}$ within the group g for an ordinal outcome $r \in \mathcal{R}^{\mathsf{Ord}}$ are

Table 6.1: The contribution of a single observation to the log-likelihood as well as the first and second derivative depending on the type of the outcome. Formulas for ordinal and categorical outcomes assume $Y = k \in \{0, \ldots, K-1\}$. General categorical outcomes have a multivariate predictor $\boldsymbol{\eta}$, the other types work with a univariate predictor η. Notation follows the one used in Section 2.2

Type	Equation	$\ell(Y \mid \boldsymbol{\eta}, \tau, \boldsymbol{c})$	$\frac{\partial}{\partial \boldsymbol{\eta}} \ell(Y \mid \boldsymbol{\eta}, \tau, \boldsymbol{c})$	$-\frac{\partial^2}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top} \ell(Y \mid \boldsymbol{\eta}, \tau, \boldsymbol{c})$
Num	(2.3)	$-\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\tau - \frac{\tau}{2}(Y-\eta)^2$	$\tau(Y-\eta)$	τ
Poi	(2.8)	$Y\eta - \exp\{\eta\}$	$Y - \exp\{\eta\}$	$\exp\{\eta\}$
Bin	(2.10)	$Y\eta - \log(1 + \exp\{\eta\})$	$Y - \mathsf{logit}^{-1}(\eta)$	$\mathsf{logit}^{-1}(\eta)\left(1 - \mathsf{logit}^{-1}(\eta)\right)$
Ord	(2.13)	$\log(q_Y) = \log(p_{Y-1} - p_Y)$	$1 - p_{Y-1} - p_Y$	$p_{Y-1}(1 - p_{Y-1}) + p_Y(1 - p_Y)$
Cat	(2.15)	$\eta_Y - \log\left(1 + \sum_{k=1}^{K-1} \exp\{\eta_k\}\right)$	$\boldsymbol{e}_Y - \mathsf{softmax}(\boldsymbol{\eta}):\ \text{if } 0 < Y \leq K-1$ $-\mathsf{softmax}(\boldsymbol{\eta}):\ \text{if } Y = 0$	$\mathsf{diag}\{\mathsf{softmax}(\boldsymbol{\eta})\} -$ $\mathsf{softmax}(\boldsymbol{\eta})\,\mathsf{softmax}(\boldsymbol{\eta})^\top$

derived:

$$\begin{aligned}
\ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{c}_r^{(g)}; \mathcal{C}\right) &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \log\left(p_{Y_{i,j}^r-1} - p_{Y_{i,j}^r}\right) \\
&\quad - \frac{1}{2}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right), \\
\frac{\partial \ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{c}_r^{(g)}; \mathcal{C}\right)}{\partial \boldsymbol{\beta}_r^{(g)}} &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \left[1 - p_{Y_{i,j}^r-1} - p_{Y_{i,j}^r}\right] \boldsymbol{x}_{i,j}^r \\
&\quad - \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right), \\
-\frac{\partial^2 \ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{c}_r^{(g)}; \mathcal{C}\right)}{\partial \boldsymbol{\beta}_r^{(g)} \partial\left(\boldsymbol{\beta}_r^{(g)}\right)^\top} &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \Big[p_{Y_{i,j}^r-1}\left(1 - p_{Y_{i,j}^r-1}\right) \\
&\quad + p_{Y_{i,j}^r}\left(1 - p_{Y_{i,j}^r}\right)\Big] \boldsymbol{x}_{i,j}^r \left(\boldsymbol{x}_{i,j}^r\right)^\top + \mathbb{D}_r^{-1}.
\end{aligned}$$

Analogously, we present these quantities for a general categorical outcome $r \in \mathcal{R}^{\mathsf{Cat}}$. For a general categorical outcome, we do not only have different $\boldsymbol{\beta}_{r,k}^{(g)}$ for each of the clusters but also for different outcome levels $k = 1, \ldots, K^r - 1$. Remind that for $k = 0$ we suppose $\boldsymbol{\beta}_{r,0}^{(g)} = \boldsymbol{0}$ by default. Notice that $\boldsymbol{\beta}_{r,k}^{(g)}$ with an arbitrary k affects the likelihood regardless of the outcome value. For that reason, full-conditional distributions of $\boldsymbol{\beta}_{r,k}^{(g)}$ are not independent between different values of k. Hence, we stack them into a long vector $\boldsymbol{\beta}_r^{(g)} = \left(\boldsymbol{\beta}_{r,1}^{(g)}, \ldots, \boldsymbol{\beta}_{r,K^r-1}^{(g)}\right)^\top$ that will be sampled at once. The log-posterior of the full-conditional distribution of $\boldsymbol{\beta}_r^{(g)}$ takes the form of:

$$\begin{aligned}
&\ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right) = \\
&\qquad \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \log\left[\eta_{i,j,Y_{i,j}^r}^{r,(g)} - \log\left(1 + \sum_{k=1}^{K^r-1} \exp\left\{\eta_{i,j,k}^{r,(g)}\right\}\right)\right] \\
&\qquad\qquad - \frac{1}{2}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right)^\top \mathbb{D}_r^{-1}\left(\boldsymbol{\beta}_r^{(g)} - \boldsymbol{\beta}_{0,r}\right).
\end{aligned}$$

The first derivative consists of the following subvectors:

$$\begin{aligned}
&\frac{\partial \ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right)}{\partial \boldsymbol{\beta}_{r,k}^{(g)}} = \\
&\qquad \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \left[\mathbb{1}_{(Y_{i,j}^r=k)} - \mathsf{softmax}_k\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right)\right] \boldsymbol{x}_{i,j}^r - \mathbb{D}_{r,k}^{-1}\left(\boldsymbol{\beta}_{r,k}^{(g)} - \boldsymbol{\beta}_{0,r,k}\right).
\end{aligned}$$

The negative Hessian matrix consists of the following blocks:

$$\begin{aligned}
-\frac{\partial^2 \ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right)}{\partial \boldsymbol{\beta}_{r,k}^{(g)} \partial\left(\boldsymbol{\beta}_{r,k}^{(g)}\right)^\top} &= \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \Big[\mathsf{softmax}_k\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right) \\
&\quad \left(1 - \mathsf{softmax}_k\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right)\right)\Big] \boldsymbol{x}_{i,j}^r \left(\boldsymbol{x}_{i,j}^r\right)^\top + \mathbb{D}_r^{-1},
\end{aligned}$$

$$-\frac{\partial^2 \ell\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r; \mathcal{C}\right)}{\partial \boldsymbol{\beta}_{r,k_1}^{(g)} \partial \left(\boldsymbol{\beta}_{r,k_2}^{(g)}\right)^\top} = \sum_{\{i:U_i=g\}} \sum_{j=1}^{n_i} \left[- \mathsf{softmax}_{k_1}\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right) \mathsf{softmax}_{k_2}\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right)\right] \boldsymbol{x}_{i,j}^r \left(\boldsymbol{x}_{i,j}^r\right)^\top,$$

where $k, k_1, k_2 \in \{1, \ldots, K^r - 1\}$ and $k_1 \neq k_2$.

If any part of the fixed effects $\boldsymbol{\beta}_r$ is common to all clusters, we need to consider the common part and the group-specific part separately. For the group-specific part the formulae are the same. Only the vector is of lower dimension because $\boldsymbol{x}_{i,j}^r$ then only contains the subset of regressors for the group-specific regression coefficients. The effects common to all clusters are sampled separately conditionally on the group-specific part, which is not part of the derivative of the predictor η in the same way as the random-effect contribution η^{R} is not included. The resulting formulae are analogous, however, they use all the units $i = 1, \ldots, n$. This feature of our implementation is discussed in more detail in Section 7.1.

6.2.3 Random effects $\boldsymbol{b}_i$

Random effects $\boldsymbol{b}_i$ are unit-specific, i.e. there is one set of random effects for each unit $i = 1, \ldots, n$. Hence, only observations belonging to unit i appear in the full-conditional distribution of $\boldsymbol{b}_i$. Each $\boldsymbol{b}_i$ consists of subvectors $\boldsymbol{b}_i^r$ for each of the outcomes $r \in \mathcal{R}$ which are modelled independently of each other given the random effects. The dependencies among the random effects $\boldsymbol{b}_i$ arise from assuming that they follow a multivariate normal distribution with general covariance matrix $\boldsymbol{\Sigma}^{(g)}$ (possibly) specific to cluster g across units. Putting all of this together similarly as in (6.4) yields the following log-posterior of the full-conditional distribution of $\boldsymbol{b}_i$:

$$\begin{aligned}
\ell\left(\boldsymbol{b}_i \,\middle|\, \mathbb{Y}_i, U_i = g, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{c}^{(g)}, \boldsymbol{\Sigma}^{(g)}; \mathcal{C}_i\right) = \text{const.} - \frac{1}{2}\boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-(g)} \boldsymbol{b}_i - \\
- \sum_{r \in \mathcal{R}^{\mathsf{Num}}} \frac{\tau_r^{(g)}}{2} \sum_{j=1}^{n_i} \left(Y_{i,j}^r - \eta_{i,j}^{r,(g)}\right)^2 + \\
+ \sum_{r \in \mathcal{R}^{\mathsf{Poi}}} \sum_{j=1}^{n_i} \left(Y_{i,j}^r \eta_{i,j}^{r,(g)} - \exp\left\{\eta_{i,j}^{r,(g)}\right\}\right) + \\
+ \sum_{r \in \mathcal{R}^{\mathsf{Bin}}} \sum_{j=1}^{n_i} \left[Y_{i,j}^r \eta_{i,j}^{r,(g)} - \log\left(1 + \exp\left\{\eta_{i,j}^{r,(g)}\right\}\right)\right] + \\
+ \sum_{r \in \mathcal{R}^{\mathsf{Ord}}} \sum_{j=1}^{n_i} \log\left(p_{Y_{i,j}^r - 1} - p_{Y_{i,j}^r}\right) + \\
+ \sum_{r \in \mathcal{R}^{\mathsf{Cat}}} \sum_{j=1}^{n_i} \log\left[\eta_{i,j,Y_{i,j}^r}^{r,(g)} - \log\left(1 + \sum_{k=1}^{K^r-1} \exp\left\{\eta_{i,j,k}^{r,(g)}\right\}\right)\right].
\end{aligned} \tag{6.11}$$

Subvectors $\boldsymbol{b}_i^r$ (or $\boldsymbol{b}_{i,k}^r$ if random effects are specific to each level of general categorical outcome r) hide within the predictor $\eta_{i,j}^{r,(g)}$ (or $\eta_{i,j,k}^{r,(g)}$ for general categorical outcome r). We use the following derivatives

$$\frac{\partial \eta}{\partial \boldsymbol{b}_i^r} = \frac{\partial\left(\eta^{\mathsf{F}} + \boldsymbol{z}^\top \boldsymbol{b}_i^r\right)}{\partial \boldsymbol{b}_i^r} = \boldsymbol{z}$$

in combination with the derivatives in Table 6.1 to compute the derivatives of full-conditional log-posterior of $\boldsymbol{b}_i$ with respect to subvectors $\boldsymbol{b}_i^r$. The explicit form of the formulae below also depends on the possibility to simplify the model by $\boldsymbol{b}_{i,1}^r = \cdots = \boldsymbol{b}_{i,K^r-1}^r$ discussed in Section 2.2.4. In case that $\boldsymbol{b}_i^r$ are common to all (except the zero level) categorical outcome values $k = 1, \ldots, K^r - 1$, the first derivative $\partial \ell \left(\boldsymbol{b}_i | \cdots \right) / \partial \boldsymbol{b}_i$ takes the following block form:

$$\begin{pmatrix} \vdots \\ \tau_r^{(g)} \sum_{j=1}^{n_i} \left(Y_{i,j}^r - \eta_{i,j}^{r,(g)}\right) \boldsymbol{z}_{i,j}^r,\ r \in \mathcal{R}^{\mathsf{Num}} \\ \vdots \\ \sum_{j=1}^{n_i} \left(Y_{i,j}^r - \exp\left\{\eta_{i,j}^{r,(g)}\right\}\right)\ r \in \mathcal{R}^{\mathsf{Poi}} \\ \vdots \\ \sum_{j=1}^{n_i} \left[Y_{i,j}^r - \mathsf{logit}^{-1}\left(\eta_{i,j}^{r,(g)}\right)\right] \boldsymbol{z}_{i,j}^r,\ r \in \mathcal{R}^{\mathsf{Bin}} \\ \vdots \\ \sum_{j=1}^{n_i} \left[1 - p_{Y_{i,j}^r - 1} - p_{Y_{i,j}^r}\right] \boldsymbol{z}_{i,j}^r,\ r \in \mathcal{R}^{\mathsf{Ord}} \\ \vdots \\ \sum_{j=1}^{n_i} \left[\mathbb{1}_{(Y_{i,j}^r \neq 0)} - \frac{\sum_{k=1}^{K^r-1} \exp\left\{\eta_{i,j,k}^{r,(g)}\right\}}{1 + \sum_{k=1}^{K^r-1} \exp\left\{\eta_{i,j,k}^{r,(g)}\right\}}\right] \boldsymbol{z}_{i,j}^r,\ r \in \mathcal{R}^{\mathsf{Cat}} \\ \vdots \end{pmatrix} - \boldsymbol{\Sigma}^{-(g)} \boldsymbol{b}_i.$$

However, if the random effects $\boldsymbol{b}_{i,k}^r$ are specific to each level $k = 1, \ldots, K^r - 1$ (the first is always zero for identifiability purposes) of the general categorical outcome we would replace the row corresponding to an outcome $r \in \mathcal{R}^{\mathsf{Cat}}$ with

$$\left(\frac{\partial \ell \left(\boldsymbol{b}_i | \cdots \right)}{\partial \boldsymbol{b}_{i,k}^r}\right)_{k=1,\ldots,K^r-1} = \left(\sum_{j=1}^{n_i} \left[\mathbb{1}_{(Y_{i,j}^r = k)} - \mathsf{softmax}_k\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right)\right] \boldsymbol{z}_{i,j}^r\right)_{k=1,\ldots,K^r-1} + \cdots,$$

where $\cdots$ on the right hand side stands for corresponding elements of $-\boldsymbol{\Sigma}^{-(g)} \boldsymbol{b}_i$ coming from the prior distribution.

With regard to the Hessian matrix, it is again better to deal with the two contributions separately. The basis of the negative Hessian matrix is formed by $\boldsymbol{\Sigma}^{-(g)}$. The other contribution comes in the form of a block-diagonal matrix, where the diagonal structure comes from the fact that $\boldsymbol{b}_i^r$ among different outcomes $r \in \mathcal{R}$ do not interact within the model specification, i.e.

$$\frac{\partial^2 \ell \left(\boldsymbol{b}_i | \cdots \right)}{\partial \boldsymbol{b}_i^{r_1} \partial \left(\boldsymbol{b}_i^{r_2}\right)^\top} = \mathbb{O}_{d_{r_1}^{\mathsf{R}} \times d_{r_2}^{\mathsf{R}}} \quad \text{for} \quad r_1, r_2 \in \mathcal{R} : r_1 \neq r_2.$$

Below is the list of diagonal blocks except for the contribution of $\boldsymbol{\Sigma}^{-(g)}$:

$$\begin{aligned}
&\tau_r^{(g)} \sum_{j=1}^{n_i} \boldsymbol{z}_{i,j}^r \left(\boldsymbol{z}_{i,j}^r\right)^\top, \quad r \in \mathcal{R}^{\mathsf{Num}},\\
&\sum_{j=1}^{n_i} \exp\left\{\eta_{i,j}^{r,(g)}\right\} \boldsymbol{z}_{i,j}^r \left(\boldsymbol{z}_{i,j}^r\right)^\top, \quad r \in \mathcal{R}^{\mathsf{Poi}},\\
&\sum_{j=1}^{n_i} \left[\mathsf{logit}^{-1}\left(\eta_{i,j}^{r,(g)}\right)\left(1-\mathsf{logit}^{-1}\left(\eta_{i,j}^{r,(g)}\right)\right)\right] \boldsymbol{z}_{i,j}^r \left(\boldsymbol{z}_{i,j}^r\right)^\top, \quad r \in \mathcal{R}^{\mathsf{Bin}},\\
&\sum_{j=1}^{n_i} \left[p_{Y_{i,j}^r-1}\left(1-p_{Y_{i,j}^r-1}\right)+p_{Y_{i,j}^r}\left(1-p_{Y_{i,j}^r}\right)\right] \boldsymbol{z}_{i,j}^r \left(\boldsymbol{z}_{i,j}^r\right)^\top, \quad r \in \mathcal{R}^{\mathsf{Ord}},\\
&\sum_{j=1}^{n_i} \left[\mathsf{softmax}_k\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right)\left(1-\mathsf{softmax}_k\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right)\right)\right] \boldsymbol{z}_{i,j}^r \left(\boldsymbol{z}_{i,j}^r\right)^\top, \quad r \in \mathcal{R}^{\mathsf{Cat}},\\
&\sum_{j=1}^{n_i} \left[-\,\mathsf{softmax}_{k_1}\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right)\mathsf{softmax}_{k_2}\left(\boldsymbol{\eta}_{i,j}^{r,(g)}\right)\right] \boldsymbol{z}_{i,j}^r \left(\boldsymbol{z}_{i,j}^r\right)^\top, \quad r \in \mathcal{R}^{\mathsf{Cat}},
\end{aligned}$$

where $k, k_1, k_2 \in \{1, \ldots, K^r - 1\}$ and $k_1 \neq k_2$. In the case when $\boldsymbol{b}_i^r$ is common to all levels $k = 1, \ldots, K^r - 1$ of a general categorical outcome $r \in \mathcal{R}^{\mathsf{Cat}}$, the corresponding block is equal to

$$\sum_{j=1}^{n_i} \frac{\sum_{k=1}^{K^r-1} \exp\left\{\eta_{i,j,k}^{r,(g)}\right\}}{\left(1+\sum_{k=1}^{K^r-1} \exp\left\{\eta_{i,j,k}^{r,(g)}\right\}\right)^2} \boldsymbol{z}_{i,j}^r \left(\boldsymbol{z}_{i,j}^r\right)^\top.$$

6.2.4 Ordered intercepts $\boldsymbol{c}$ for ordinal outcomes

The prior distribution for parameter $\boldsymbol{c}$ is specified through the probabilities $\boldsymbol{\pi}$ by (4.14). Specifying the prior for the probabilities allows for a more straightforward inclusion of prior knowledge. Both $\boldsymbol{c}$ and $\boldsymbol{\pi}$ cannot directly be used in combination with a Metropolis proposal without taking into account the limitation of the corresponding parametric space. Hence, we express $\boldsymbol{\pi}$ in terms of the new parameter $\boldsymbol{a}$ in the following way:

$$\pi_k = \mathsf{softmax}_k(\boldsymbol{a}) := \frac{e^{a_k}}{\sum_{k'=0}^{K-1} e^{a_{k'}}}$$

with $a_k = \log(\pi_k/\pi_0)$, $k = 1, \ldots, K-1$ and $a_0 = 0$, thus implying

$$\begin{aligned}
c_k &\equiv c_k(\boldsymbol{a}) = \log \frac{1+e^{a_1}+\cdots+e^{a_k}}{e^{a_{k+1}}+\cdots+e^{a_{K-1}}}, && k = 0, \ldots, K-2,\\
a_k &\equiv a_k(\boldsymbol{c}) = \log \frac{\mathsf{logit}^{-1}(c_k) - \mathsf{logit}^{-1}(c_{k-1})}{\mathsf{logit}^{-1}(c_0)}, && k = 1, \ldots, K-1.
\end{aligned}$$

Note that we dropped the ordinal outcome index $r \in \mathcal{R}^{\mathsf{Ord}}$ and the superscript (g) for the g-th cluster for simplicity. The prior (4.15) over $\boldsymbol{\pi}$ translates to the following form in terms of $\boldsymbol{a}$:

$$p(\boldsymbol{a}) \propto \prod_{k=0}^{K-1} (\pi_k)^{\alpha_k - 1} \cdot \prod_{k=0}^{K-1} \frac{e^{a_k}}{\sum_{k'=1}^{K} e^{a_{k'}}} = \prod_{k=0}^{K-1} \left(\mathsf{softmax}_k(\boldsymbol{a})\right)^{\alpha_k}.$$

The logarithm of this can be easily differentiated:

$$\begin{aligned}
\log p\left(\boldsymbol{a}\right) &= \sum_{k=0}^{K-1} \alpha_k a_k - (\alpha_0 + \cdots + \alpha_{K-1}) \log\left(1 + \sum_{k=1}^{K-1} \exp\{a_k\}\right), \\
\frac{\partial \log p\left(\boldsymbol{a}\right)}{\partial \boldsymbol{a}} &= \boldsymbol{\alpha} - (\alpha_0 + \cdots + \alpha_{K-1})\, \mathsf{softmax}(\boldsymbol{a}), \\
-\frac{\partial^2 \log p\left(\boldsymbol{a}\right)}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\top} &= (\alpha_0 + \cdots + \alpha_{K-1}) \\
&\qquad \Big(\mathsf{diag}\{\mathsf{softmax}(\boldsymbol{a})\} - \mathsf{softmax}(\boldsymbol{a})\, \mathsf{softmax}(\boldsymbol{a})^\top\Big).
\end{aligned}$$

The parameter vector $\boldsymbol{a} = (a_1, \ldots, a_{K-1})^\top \in \mathbb{R}^{K-1}$ is not restricted. Hence, we can propose a new value for $\boldsymbol{a}$ using a usual Metropolis proposal step and obtain $\boldsymbol{c}$ or $\boldsymbol{\pi}$ using the backward transformation described above. The log-posterior of the full-conditional distribution of $\boldsymbol{a}_r^{(g)}$ takes the following form:

$$\begin{aligned}
\ell\left(\boldsymbol{a}_r^{(g)} \,\middle|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{\beta}_r^{(g)}; \mathcal{C}^r\right) &= \text{const.} + \log p\left(\boldsymbol{a}_r^{(g)}\right) + \\
+ \sum_{\{i : U_i = g\}} \sum_{j=1}^{n_i} \log \Bigg[& \underbrace{\mathsf{logit}^{-1}\left(\eta_{i,j}^{r,(g)} - c_{Y_{i,j}^r - 1}\left(\boldsymbol{a}_r^{(g)}\right)\right)}_{p_{Y_{i,j}^r - 1}\left(\boldsymbol{a}_r^{(g)}\right)} - \underbrace{\mathsf{logit}^{-1}\left(\eta_{i,j}^{r,(g)} - c_{Y_{i,j}^r}\left(\boldsymbol{a}_r^{(g)}\right)\right)}_{p_{Y_{i,j}^r}\left(\boldsymbol{a}_r^{(g)}\right)} \Bigg],
\end{aligned}$$

where we use the notation from Section 2.2.3 enriched by highlighted dependence on $\boldsymbol{a}_r^{(g)}$. Focusing on outcome $r \in \mathcal{R}^{\mathsf{Ord}}$ and group g, we strip away the nuisance indices to obtain the expression

$$\begin{aligned}
\ell\left(\boldsymbol{a} \,|\, \cdots\right) = \text{const.} + \log p\left(\boldsymbol{a}\right) + & \sum_{\{i : U_i = g\}} \sum_{j=1}^{n_i} \sum_{k=0}^{K-1} \mathbb{1}_{(Y_{i,j} = k)} \cdot \\
& \log \Bigg[\underbrace{\mathsf{logit}^{-1}\left(\eta_{i,j} - \log \frac{1 + e^{a_1} + \cdots + e^{a_{k-1}}}{e^{a_k} + \cdots + e^{a_{K-1}}}\right)}_{p_{k-1}} - \\
& \qquad - \underbrace{\mathsf{logit}^{-1}\left(\eta_{i,j} - \log \frac{1 + e^{a_1} + \cdots + e^{a_k}}{e^{a_{k+1}} + \cdots + e^{a_{K-1}}}\right)}_{p_k} \Bigg].
\end{aligned}$$

Before we obtain its derivatives, we first present the derivatives of the probabilities p_k and $q_k = p_{k-1} - p_k$ (remember also $1 = p_{-1} > p_0 > p_1 > \cdots > p_{K-1} = 0$) with respect to a_l:

$$\begin{aligned}
\frac{\partial p_{k_1}}{\partial c_{k_2}} &= \frac{\partial\, \mathsf{logit}^{-1}(\eta - c_{k_1})}{\partial c_{k_2}} \\
&= \begin{cases} -p_k(1 - p_k), & \text{if } k = k_1 = k_2 = 0, \ldots, K-2, \\ 0, & \text{otherwise}, \end{cases} \\
\frac{\partial c_k}{\partial a_l} &= \begin{cases} \dfrac{e^{a_l}}{1 + e^{a_1} + \cdots + e^{a_k}}, & \text{if } 1 \leq l \leq k \leq K-2, \\ \dfrac{-e^{a_l}}{e^{a_{k+1}} + \cdots + e^{a_{K-1}}}, & \text{if } 0 \leq k < l \leq K-1, \\ 0, & \text{otherwise}, \end{cases} \\
\frac{\partial \log(p_{k-1} - p_k)}{\partial a_l} &= -\frac{1}{p_{k-1} - p_k} \left[p_{k-1}(1 - p_{k-1}) \frac{\partial c_{k-1}}{\partial a_l} - p_k(1 - p_k) \frac{\partial c_k}{\partial a_l} \right],
\end{aligned}$$

for $k = 0, \ldots, K-2$, and $l = 1, \ldots, K-1$.

Finally, we can evaluate the gradient of the log-posterior of the full-conditional distribution of the parameter $\boldsymbol{a}_r^{(g)}$ by

$$\begin{aligned}\frac{\partial \ell\left(\boldsymbol{a} \mid \cdots\right)}{\partial \boldsymbol{a}} = \frac{\partial \log p\left(\boldsymbol{a}\right)}{\partial \boldsymbol{a}} - \\ \sum_{\{i: U_i = g\}} \sum_{j=1}^{n_i} \sum_{k=0}^{K-1} \mathbb{1}_{(Y_{i,j}=k)} \frac{p_{k-1}(1-p_{k-1})\dfrac{\partial c_{k-1}}{\partial \boldsymbol{a}} - p_k(1-p_k)\dfrac{\partial c_k}{\partial \boldsymbol{a}}}{p_{k-1} - p_k}\end{aligned}$$

for a specific choice of ordinal outcome $r \in \mathcal{R}^{\mathsf{Ord}}$ and group g.

Next, we determine the second derivatives of c_k with respect to a_{l_1} and a_{l_2} for $1 \leq l_1 \leq l_2 \leq K-1$

$$\frac{\partial^2 c_k}{\partial a_{l_1} \partial a_{l_2}} = \begin{cases} \dfrac{\partial c_k}{\partial a_l}\left(1 - \dfrac{\partial c_k}{\partial a_l}\right) & \text{if } 1 \leq l = l_1 = l_2 \leq k \leq K-2, \\ \dfrac{\partial c_k}{\partial a_l}\left(1 + \dfrac{\partial c_k}{\partial a_l}\right) & \text{if } 0 \leq k < l = l_1 = l_2 \leq K-1, \\ -\dfrac{\partial c_k}{\partial a_{l_1}} \dfrac{\partial c_k}{\partial a_{l_2}} & \text{if } 1 \leq l_1 < l_2 \leq k \leq K-2, \\ -\dfrac{\partial c_k}{\partial a_{l_1}} \dfrac{\partial c_k}{\partial a_{l_2}} & \text{if } 0 \leq k < l_1 < l_2 \leq K-1, \\ 0 & \text{otherwise.} \end{cases}$$

Now we can proceed with the second derivatives of individual model contributions for $k = 0, \ldots, K-1$ and $l_1, l_2 = 1, \ldots, K-1$.

$$\begin{aligned} -\frac{\partial^2 \log(p_{k-1} - p_k)}{\partial a_{l_1} \partial a_{l_2}} = & \frac{\partial^2 c_{k-1}}{\partial a_{l_1} \partial a_{l_2}} \frac{p_{k-1}(1-p_{k-1})}{p_{k-1} - p_k} - \\ & - \frac{\partial^2 c_k}{\partial a_{l_1} \partial a_{l_2}} \frac{p_k(1-p_k)}{p_{k-1} - p_k} + \\ & + \frac{\partial c_{k-1}}{\partial a_{l_1}} \frac{\partial c_{k-1}}{\partial a_{l_2}} \frac{p_{k-1}(1-p_{k-1})[p_{k-1}^2 + p_k(1-2p_{k-1})]}{(p_{k-1} - p_k)^2} - \\ & - \frac{\partial c_{k-1}}{\partial a_{l_1}} \frac{\partial c_k}{\partial a_{l_2}} \frac{p_{k-1}(1-p_{k-1})p_k(1-p_k)}{(p_{k-1} - p_k)^2} + \\ & + \frac{\partial c_k}{\partial a_{l_1}} \frac{\partial c_k}{\partial a_{l_2}} \frac{p_k(1-p_k)[p_k^2 + p_{k-1}(1-2p_k)]}{(p_{k-1} - p_k)^2} - \\ & - \frac{\partial c_k}{\partial a_{l_1}} \frac{\partial c_{k-1}}{\partial a_{l_2}} \frac{p_{k-1}(1-p_{k-1})p_k(1-p_k)}{(p_{k-1} - p_k)^2}. \end{aligned}$$

Finally, we can express the negative Hessian matrix of the log-posterior of the full-conditional distribution of the parameter $\boldsymbol{a}_r^{(g)}$ in the following way:

$$-\frac{\partial^2 \ell\left(\boldsymbol{a} \mid \cdots\right)}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\top} = -\frac{\partial^2 \log p\left(\boldsymbol{a}\right)}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\top} - \sum_{\{i: U_i = g\}} \sum_{j=1}^{n_i} \sum_{k=0}^{K-1} \mathbb{1}_{(Y_{i,j}=k)} \frac{\partial^2 \log(p_{k-1} - p_k)}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\top}$$

for a specific choice of ordinal outcome $r \in \mathcal{R}^{\mathsf{Ord}}$ and group g.

6.3 Metropolis within Gibbs sampling

As declared above, we aim to estimate the posterior $p\left(\boldsymbol{\Psi}\,\middle|\,\mathbb{Y}^{\mathsf{obs}};\mathcal{C}\right)$ of the *GLMM-based* model by Gibbs sampling where some of the steps have to be replaced by a Metropolis proposal step. We have already derived the needed full-conditional distributions (Sections 5.1 and 6.1) and explained how new values for the rest of the parameters will be proposed for acceptance (Section 6.2). All these steps are combined into Algorithm 2.

Similarly as for Algorithm 1, we have to set the length of the burn-in period B, the length of the chain M, thinning (to reduce autocorrelation), number of chains to be sampled. Then, alongside the fixed hyperparameter values $\mathcal{H}_0$ there is a large number of *setting* or *tuning* parameters. Namely,

- whether also random effects for general categorical outcomes $r \in \mathcal{R}^{\mathsf{Cat}}$ should be level-specific or not (discussed in Sections 2.2.4 and 6.2.3),
- how often should be the proposal distribution updated,
- how many proposals should be performed in one step,
- tolerance and maximal number of iterations for Newton-Raphson method,
- multiplicative constants $c_{\boldsymbol{\omega}}$ for $\boldsymbol{\omega} \in \left\{e_0^{\star}, \boldsymbol{\beta}_r^{(g)}, \boldsymbol{b}_i, \boldsymbol{a}_r^{(g)}\right\}$ to control the size of incremental steps of Metropolis proposals.

Regarding the frequency of proposals, each proposed parameter should have its own counter since the last update of the proposal distribution which should be compared to the tuned frequency prior the update. Exceptionally, the update of the proposal distribution can be performed sooner, e.g. when unit i changes the assigned cluster and, consequently, the full-conditional distribution of $\boldsymbol{b}_i$ is significantly changed, which requires an appropriate reaction from the proposal distribution.

Important role in this model is played by the *sparse finite mixture* framework by Malsiner-Walli et al. (2016) which is used to estimate the number of underlying components by the number G_+ of non-empty components and inducing a prior heavily favouring sparse partitions. First, the maximal number G_{max} of mixture components has to be set. Then, each group-specific parameter requires mutually different values for each cluster. For that reason, we again generate initial partition by assigning cluster labels to all units uniformly.

Analogously to Algorithm 1, we set the initial values of unknown parameters according to estimates coming from univariate models ignoring the random effects. Parameters for numeric, count and binary outcomes are initialized by the estimates coming from the standard `glm` function in base R (R Core Team, 2022). Ordinal logistic regression is estimated via `polr` from `MASS` (Venables and Ripley, 2002) where `zeta` coincides with our $\boldsymbol{c}_r^{(g)}$. Multinomial logistic regression for general categorical outcomes is performed by `multinom` from `nnet` (Venables and Ripley, 2002). Missing values $\mathbb{Y}^{\mathsf{mis}}$ and $\boldsymbol{\Sigma}$ do not have to be initialized since their value is not needed in any sampling step above within the for cycle of Algorithm 2. This time the random effects $\boldsymbol{b}_i$ are rather sampled from the normal distribution than estimated by appropriate mixed-models.

Algorithm 2 Metropolis within Gibbs sampling for the *GLMM-based* model

```
  Input: Data $\mathbb{Y}$ of longitudinal profiles of $n$ units and covariates $\mathcal{C}$.
  Set $B$, $M$ and other tuning parameters.
  Choose the maximal number of clusters $G_{\mathsf{max}}$ and fix the hyperparameters $\mathcal{H}_0$.
  Declare or find initial values $\boldsymbol{\Psi}^0$ in this order:
    • divide $n$ units randomly into $G_{\mathsf{max}}$ clusters by $U_i^0 \sim \mathsf{Unif}\,\{1, \ldots, G_{\mathsf{max}}\}$;
    • $e_0^0 := a_e/b_e$;
    • estimate $\boldsymbol{\beta}_r^{(g),0}$, $\tau_r^{(g),0}$ and $\boldsymbol{c}_r^{(g),0}$ using linear regression, Poisson log-linear model, logistic regression, ordinal regression and multinomial logistic regression ignoring any potential random effects and using available data within cluster $g$;
    • randomly generate $b_{i,j}^0$ from $\mathsf{N}\,(0, \sigma^2)$ with $\sigma^2 > 0$ low, e.g. $\sigma^2 = 10^{-4}$;
    • randomly generate $\boldsymbol{\Sigma}^0$ as diagonal matrices with Gamma-distributed diagonal elements (with large variance);
    • estimate $\mathbb{Q}^0$ by inverting the mean of all $\boldsymbol{\Sigma}^{(g),0}$ divided by $\nu_0$.

Metropolis within Gibbs sampling  Always use the very last known values of other parameters, i.e. either from $\boldsymbol{\Psi}^{m-1}$ or $\boldsymbol{\Psi}^m$:
  for $m$ in $1 : (B + M)$ do
    if $Y_{i,j}^r$ is missing then
      • $Y_{i,j}^{r,m} \Big| U_i = g, \boldsymbol{b}_i^r, \boldsymbol{\beta}_r^{(g)}, \tau_r^{(g)}, \boldsymbol{c}_r^{(g)}; \mathcal{C}_{i,j} \overset{(6.2)}{\sim} p_{\mathsf{t}(r)}\left(Y_{i,j}^r \Big| \boldsymbol{b}_i^r, \boldsymbol{\beta}_r^{(g)}, \tau_r^{(g)}, \boldsymbol{c}_r^{(g)}; \mathcal{C}_{i,j}\right)$;
    end if
    • $\boldsymbol{\beta}_r^{(g),m} \Big| \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \tau_r^{(g)}, \boldsymbol{\beta}_{0,r}, \mathbb{D}_r; \mathcal{C} \overset{(5.7)}{\sim}$
        $\mathsf{N}_{d_r^{\mathsf{F}}}\left(\widetilde{\boldsymbol{\beta}}_r^{(g)}, \frac{1}{\tau_r^{(g)}}\left[\left(\mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r\right)^\top \mathbb{X}_{\mathcal{N}_g(\boldsymbol{U})}^r + \mathbb{D}_r^{-1}\right]^{-1}\right) \quad$ for $r \in \mathcal{R}^{\mathsf{Num}}$;
    • $\tau_r^{(g),m} \Big| \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{\beta}_r^{(g)}, \boldsymbol{\beta}_{0,r}, \mathbb{D}_r, a_\tau, b_\tau; \mathcal{C} \overset{(5.6)}{\sim} \Gamma\left(\widetilde{a}_{\tau,r}^{(g)}, \widetilde{b}_{\tau,r}^{(g)}\right) \quad$ for $r \in \mathcal{R}^{\mathsf{Num}}$;
    • update proposal distribution (Alg. 6) for $\boldsymbol{\beta}_r^{(g)}, r \in \mathcal{R} \setminus \mathcal{R}^{\mathsf{Num}}$ according to Sect. 6.2.2;
    • propose and accept/deny $\boldsymbol{\beta}_r^{(g),m}, r \in \mathcal{R} \setminus \mathcal{R}^{\mathsf{Num}}$ according to Sect. 6.2;
    • update proposal distribution (Alg. 6) for $\boldsymbol{a}_r^{(g)}, r \in \mathcal{R}^{\mathsf{Ord}}$ according to Sect. 6.2.4;
    • propose and accept/deny $\boldsymbol{a}_r^{(g),m}, r \in \mathcal{R}^{\mathsf{Ord}}$ according to Sect. 6.2;
    • transform $\boldsymbol{a}_r^{(g),m}, r \in \mathcal{R}^{\mathsf{Ord}}$ into $\boldsymbol{c}_r^{(g),m} = c\left(\boldsymbol{a}_r^{(g),m}\right)$;
    • $\boldsymbol{\Sigma}^{-(g),m} \,\Big|\, \boldsymbol{U}, \boldsymbol{b}, \mathbb{Q}, \nu_0 \overset{(5.9)}{\sim} \mathsf{W}_{d^{\mathsf{R}}}\left(\widetilde{\mathbb{Q}}^{(g)}, n^{(g)}(\boldsymbol{U}) + \nu_0\right)$;
    • $(\mathbb{Q}^m)^{-1} \,\Big|\, \boldsymbol{\Sigma}, \nu_0, \nu_1, \mathbb{D}_{\mathbb{Q}} \overset{(5.8)}{\sim} \mathsf{W}_{d^{\mathsf{R}}}\left(\left[\sum_{g=1}^{G_{\mathsf{max}}} \boldsymbol{\Sigma}^{-(g)} + \mathbb{D}_{\mathbb{Q}}^{-1}\right]^{-1}, G_{\mathsf{max}}\nu_0 + \nu_1\right)$;
    • update proposal distribution (Alg. 6) for $\boldsymbol{b}_i$ according to Sect. 6.2.3;
    • propose and accept/deny $\boldsymbol{b}_i^m$ according to Sect. 6.2;
    • $\boldsymbol{w}^m \,|\, \boldsymbol{U}, e_0 \overset{(5.2)}{\sim} \mathsf{Dir}_{G_{\mathsf{max}}}\left(\boldsymbol{n}(\boldsymbol{U}) + e_0\mathbf{1}\right)$;
    • $U_i^m \;\sim\; \mathsf{P}\left[U_i = g \,|\, \mathbb{Y}_i, \boldsymbol{b}_i, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{c}, \boldsymbol{\Sigma}; \mathcal{C}_i\right] \overset{(6.3)}{=} \cdots$;
    • update proposal distribution (Alg. 6) for $e_0^\star$ according to Sect. 6.2.1;
    • propose and accept/deny $e_0^{\star,m}$ according to Sect. 6.2;
    • transform $e_0^{\star,m}$ into $e_0^m = \exp\{e_0^{\star,m}\}$.
  end for
```

The order of parameters sampled within Algorithm 2 is somewhat changed compared to Algorithm 1, the structure of the algorithm is inspired by Frühwirth-Schnatter and Malsiner-Walli (2019). However, the Metropolis within Gibbs sampler would work under any order. The only benefit one could get from a specific order is smoother transition from the initial values to the first sampled state. Gibbs sampler could be even improved by arbitrarily permuting the order of parameters for each m.

Now consider the group-specific parameters, e.g. $\boldsymbol{\beta}_r^{(g)}$, $\tau_r^{(g)}$, $\boldsymbol{a}_r^{(g)}$, $\boldsymbol{\Sigma}^{(g)}$. They require only data from units currently in the group g, which means that whenever the cluster is empty ($\mathcal{N}_g(\boldsymbol{U}) = \emptyset$), the full-conditional distribution consists only of the assumed prior. We still use Metropolis proposal steps in such situations, although the new value could be sampled from the prior directly. Since the prior distribution serves as the only source of information for the empty clusters, its strength is very important with respect to the sparse finite mixture modelling as it regulates the willingness of becoming en empty cluster. Informative prior with low variance results in high penalization, which shrinks the estimates towards the prior mean. Then, it allows only strong signals from the data to be detected, hence, negligible signals are overlooked and sparsity is achieved. However, in applications one has to balance out the level of regularization with respect to the sample size to not to overshrink.

Algorithm 3 Post-processing the output of MCMC under sparse finite mixture

```
Input: Sampled Markov chain $\{\boldsymbol{\Psi}^1, \dots, \boldsymbol{\Psi}^M\}$ using Algorithm 2.
for $m$ in $1 : M$ do
    • $n_g^m := n^{(g)}(\boldsymbol{U}^m) = \sum_{i=1}^{n} \mathbb{1}_{(U_i^m = g)}$;
    • $G_+^m := G_{\mathsf{max}} - \sum_{g=1}^{G_{\mathsf{max}}} \mathbb{1}_{(n_g^m = 0)}$.
end for
• Estimate the number of non-empty clusters by
      $$\widehat{G}_+ = \underset{g=1,\dots,G}{\arg\max} \sum_{m=1}^{M} \mathbb{1}_{(G_+^m = g)}.$$
• Create a subset of posterior draws $m = 1, \dots, M$ such that $G_+^m = \widehat{G}_+$,
  denote the number of such draws by $\widehat{M}$.
• Apply $k$-means clustering (Algorithm 7) with $\widehat{G}_+$ to dataset (of length $\widehat{G}_+ \cdot \widehat{M}$)
  consisting of draws of cluster-specific parameters within $\boldsymbol{\theta}$ that correspond
  to non-empty clusters stacked under each other.
• Take all $\widehat{G}_+$ obtained classification indices for $m$-th draw and check whether
  it results in a permutation of $\{1, \dots, \widehat{G}_+\}$.
• Count the number $M_\rho$ of draws that do *not* lead to a permutation. If it is
  large, avoid any inference that is sensitive to label switching.
• For all remaining $\widehat{M} - M_\rho$ MCMC draws, obtain a unique labelling by
  reordering each of the draws by the corresponding permutation.
```

When the sparsity is achieved, there is still a question of the final inference. Especially, estimation of the number of clusters and inference regarding the cluster-specific parameters has to be performed with caution since the number of non-empty clusters may differ among different draws from the posterior. For each draw m, the cluster indicators $\boldsymbol{U}^m$ induce cluster occupation numbers $\boldsymbol{n}^m = (n_1^m, \ldots, n_G^m)^\top$ and a specific number of non-empty components $G_+^m = G - \sum_{g=1}^{G} \mathbb{1}_{(n_g^m = 0)}$.

We estimate the number of data clusters as suggested by Malsiner-Walli et al. (2016). They use the mode $\widehat{G}_+$ of the posterior of the number of filled components as an estimator for the number of clusters in the data:

$$\widehat{G}_+ = \operatorname*{arg\,max}_{g \in \{1, \ldots, G\}} \sum_{m=1}^{M} \mathbb{1}_{(G_+^m = g)}. \tag{6.12}$$

Then, for the subsequent inference only those MCMC draws are considered where the number of filled components coincides exactly with the mode $\widehat{G}_+$. The MCMC draws where a different number of components is filled are discarded and omitted from the further analysis, see Algorithm 3 for details.

Moreover, before group-specific inference can be performed based on the MCMC samples, one potentially needs to resolve label switching, see Section 4.3. Since the posterior is multi-modal with modes corresponding to all parameterisations obtained by permuting the labels of unique components, the component labels may be switched across different draws of the MCMC sampler and a unique labelling needs to be obtained to determine an identified model where group-specific inference is possible. We suggest to use the procedure proposed in Frühwirth-Schnatter (2011) and Malsiner-Walli et al. (2016) to resolve label-switching with the later describing a method applicable when pursuing the sparse finite mixture approach. The steps are also covered by Algorithm 3.

In our simulation study and the applications, we observed that the number of filled components usually stabilises during MCMC sampling at a specific number, usually representing the lower bound of data clusters required to provide an adequate fit for the data. Initialising using a partition with all components being filled, we noted that during the first iterations of the MCMC algorithm superfluous components are emptied and only the necessary number of components required to represent the group structure in the data set remains filled. Monitoring thus the number of filled components serves as a means to assess convergence of the MCMC chain and thus decide on a suitable number of burn-in iterations to discard.

We also noted that label switching did not occur during MCMC sampling after the burn-in samples are omitted in our simulation study and the applications. Using a multivariate regression model with repeated measurements for units and avoiding redundant mixture components induces rather crisp classifying probabilities. They induce well separated modes and prevent the sampler also to move between these modes. Hence, for these analyses there was no need to apply a procedure for resolving label switching and assigning suitable labels to components such that they correspond to an identified model.

6.4 Simulation study

We performed a simulation study to demonstrate the performance of the *GLMM-based* model under various settings. We were particularly interested in assessing how the structure of the sampled data as well as the data generating process affects

1) the ability to estimate the number of data clusters,

2) the clustering performance measured by the misclassification rates,

3) the accuracy of the model parameter estimates.

Simulation design

A wide range of parameters are selected to specify the simulation study. Some parameters vary across the settings to study their impact on performance, while others are kept fixed. In particular, the sample size is varied with values $n \in \{100, 250, 500, 1000\}$ and the number of true data clusters $G \in \{2, 3\}$. Regarding the panel structure, we use a rather challenging setting of only $n_i = 4$ observations per unit in order to mimic the panel structure of the EU-SILC dataset.

For each data set we generate one outcome per type – numeric Y^{N}, binary Y^{B}, ordinal Y^{O} with $K^{\mathsf{O}} = 5$ levels and general categorical Y^{C} with $K^{\mathsf{C}} = 4$ levels. Unfortunately, the implementation of the method did not yet include the possibility for a count-type outcome by the time the simulation study was performed. Hence, count outcome is not considered, which will also be the case in Chapter 8 where the EU-SILC dataset will be analysed. With respect to the random-effects part, we only consider a random intercept term for each type of outcome $\boldsymbol{b}_i = (b_i^{\mathsf{N}}, b_i^{\mathsf{B}}, b_i^{\mathsf{O}}, b_i^{\mathsf{C}})^\top \sim \mathsf{N}_4\left(\boldsymbol{0},\, \boldsymbol{\Sigma}\right)$ and assume that the covariance matrix $\boldsymbol{\Sigma}$ of the random effects is the same across clusters and may be decomposed into standard deviations and correlation matrix such that

$$\boldsymbol{\Sigma} = \boldsymbol{S} \begin{pmatrix} 1 & -0.5 & -0.5 & -0.4 \\ -0.5 & 1 & 0.3 & 0.4 \\ -0.5 & 0.3 & 1 & 0.2 \\ -0.4 & 0.4 & 0.2 & 1 \end{pmatrix} \boldsymbol{S}.$$

with $\boldsymbol{S} = \mathsf{diag}\{0.5, 0.5, 0.5, 0.5\}$. A common random-effects structure is then also used when fitting the model.

The fixed-effects part of the predictor consists of an intercept term and one other covariate $x \in (0, 1)$. This covariate represents time and is sampled in such a way that the values are close to each other for the same unit. In particular, we use the simulation parameter $\xi = \frac{1}{3}$ to define the length of the observational window for one unit, i.e. for each unit only a third of the total length of the interval is admissible for values of x. To obtain the x values for each unit i, first, the centre of the interval is sampled by $x_{c,i} \sim \frac{\xi}{2} \cdot \mathsf{Unif}\left\{1, \ldots, \frac{2}{\xi} - 1\right\}$ and then n_i values for unit i are sampled from $\mathsf{Unif}\left(x_{c,i} - \frac{\xi}{2}, x_{c,i} + \frac{\xi}{2}\right)$ and ordered. Marginally, for $\xi < 1$ the distribution of x is not $\mathsf{Unif}\,(0, 1)$ since the intervals at the boundary $\left(0, \frac{\xi}{2}\right)$ and $\left(1 - \frac{\xi}{2}, 1\right)$ have lower probability. Note that this setting is selected to resemble the structure of the rotational panel in the EU-SILC data set.

We explore several different ways how the time covariate affects the outcome:

a) no effect of time at all (`no`),

b) a slope term common to all clusters (`parallel`),

c) different intercepts and slopes in each cluster resulting in a crossing (`cross`).

We follow the same scheme when specifying the models for estimation, considering models where no time effect is included, a common slope for time and a group-specific slope for time. Examples of the predictors simulated for the different time parametrizations and number of clusters G are illustrated in Figure 6.1.

The implementation of the MCMC sampler for the *GLMM-based* model allows us to choose which of the fixed effects $\beta_{r,j}$ will be group-specific and which not, see Section 7.1 for more details. The intercept term is always (both when generating the data set and when estimating) considered to be group-specific. This ensures some differences between clusters. While presence and group-specificity for the effect of time will create different scenarios under which the model is estimated.

The numerical outcome is obtained by adding an error term with group-specific standard deviation, $\{0.5, 0.8\}$ for $G = 2$ and $\{0.5, 0.75, 1\}$ for $G = 3$, to the linear predictor. For the ordinal outcome, group-specific equidistant ordered intercepts are used (i.e. typically whole numbers shifted by a certain constant amount to have reasonable frequencies of outcome values in each cluster). Three different specifications of intercepts (e.g. using an exchange of monotonicity type) are required to obtain the predictors for the categorical outcome with $K^{\mathsf{C}} = 4$ levels.

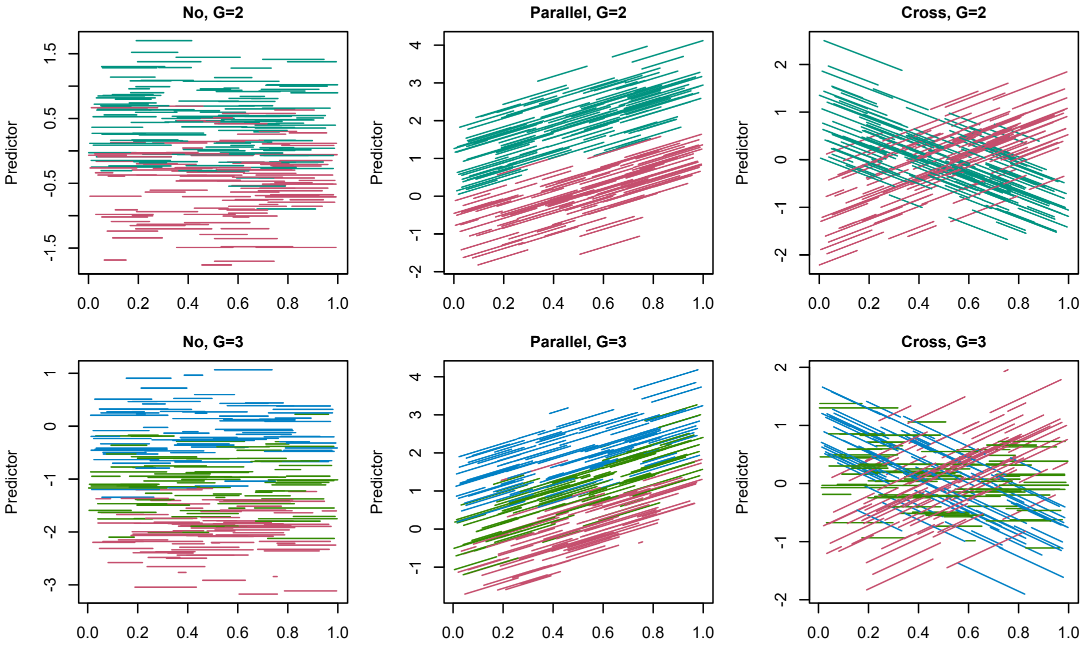


Figure 6.1: Lines connecting predictors of $n = 250$ individual units generated from G clusters for different types of time effects. The maximum length of the observational window is $\xi = \frac{1}{3}$.

We generate 200 data sets for each considered data setting. For Bayesian inference, the prior distributions together with their parameter values are specified as outlined in Section 4.2. For estimating the number of data clusters or assessing the clustering abilities, we initialise the Markov chain with the maximal number of components $G_{\mathsf{max}} = 10$ considered for the mixture model. A burn-in period of $B = 500$ samples was enough to then use the next $M = 10\,000$ sampled parameter and latent variable values to approximate their posterior distributions. Subjects were classified using the sampled indicators U_i, leaving units unclassified when less than 60% of these indicators assigned the unit to the same cluster.

Estimating number of data clusters and classifying units

In the following we assess the ability of the proposed approach to estimate the number of data clusters and evaluate the classification performance, focusing in particular on the benefit incurred through joint modelling of the outcome variables. We consider the `cross` parametrization of time with $\xi = \frac{1}{3}$ for data generation and also use a suitable model specification for estimation to be able to capture these effects. We estimate the model for each type of outcome separately as well as all four outcomes of different types jointly.

Results indicate that the performance regarding the estimation of the number of data clusters G_+ is rather comparable regardless of the type of outcome used and also when all outcomes are modelled jointly, see Figure 6.2 depicting histograms of the generated number of non-empty clusters G_+^m. Sample size had an effect with only one or two data clusters being selected for $n = 100$ regardless of if the true number of data clusters is 2 or 3. For $G = 2$ and $n = 250$ the number of data clusters was in general already correctly identified, whereas $n = 500$ was required for $G = 3$ to achieve a good performance.

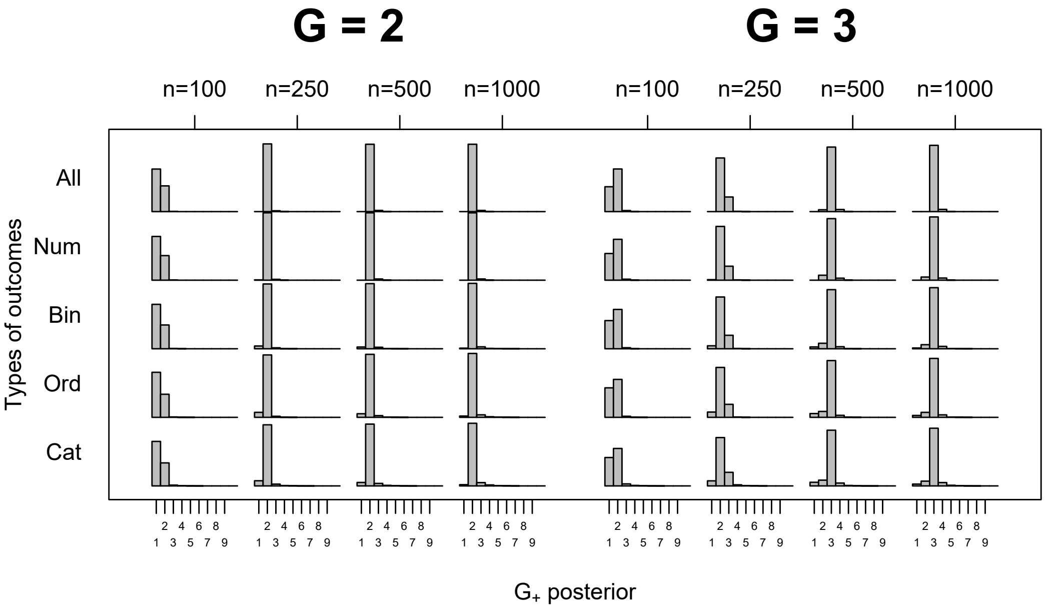


Figure 6.2: Histograms of the draws G_+^m across all 200 replications of the dataset of sample size n under `cross` effect of time. G is the true number of underlying clusters.

Once the number of clusters has been estimated by $\widehat{G}_+$ (6.12), we proceed with classification according to point estimates of $u_{i,g}$ based on sampled allocation indicators U_i^m by the rule (P1) with threshold 0.6, see Classification probabilities subsection of Section 4.3. Figure 6.3 provides an overview on the proportions of correctly classified, unclassified and misclassified units when using either only a single outcome variable or using all four outcome variables jointly. Similarly as in Figure 6.2, the results for the single outcome variables are shown in the rows labelled "Num" for numeric outcome, "Bin" for binary outcome, "Ord" for ordinal outcome and "Cat" for general categorical outcome. The results when modelling all four outcomes jointly are shown on top in the row labelled "All". In addition the sample size n and the true number of data clusters are also varied.

Figure 6.3 clearly shows a general pattern of an increase in sample size n improving the classification performance. This certainly also is partly due to the underestimation of G_+ for $n \in \{100, 250\}$. In case the number of data clusters is underestimated, a high misclassification rate naturally results. Also the classification performance is in general better if the true number of data clusters is 2 instead of 3.

Figure 6.3 also highlights the impact of the type of outcome on the classification performance. If only a single outcome is considered, the numeric outcome performs best, while the single categorical outcome classifies barely better than a completely random classification. Modelling all types together clearly outperforms the single models and achieves the highest correct classification rates indicating the advantage of using a modelling approach which allows to jointly model the data.

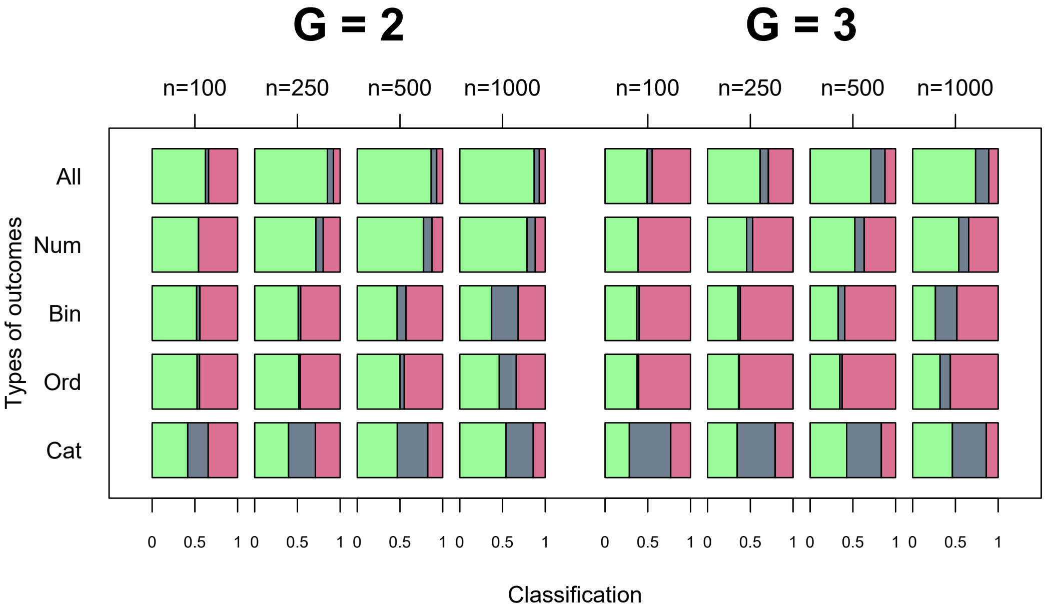


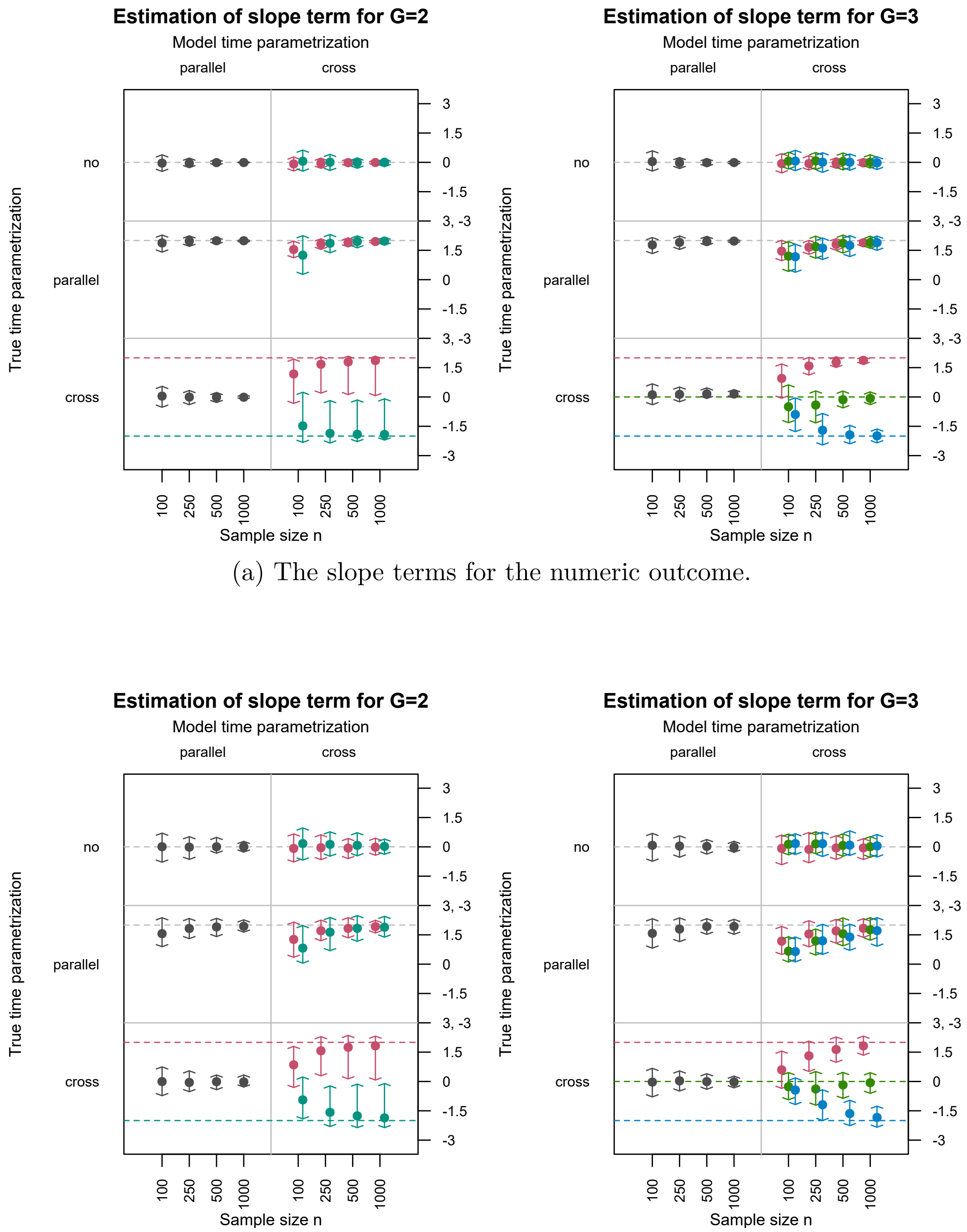
Figure 6.3: Proportions of correctly classified (green), unclassified (grey) and misclassified (red) units in dependence of the types of outcomes used, sample size n and the true number of data clusters G. The number of data clusters used for classification are estimated based on $\widehat{G}_+$, the most frequent number of non-empty components during MCMC sampling with $G_{\mathsf{max}} = 10$.

Estimating model parameters

Regarding the accuracy of the model parameter estimates, we focus on the estimation of the fixed effects $\boldsymbol{\beta}$. In many applications these parameters will be of core interest for characterising the clusters identified and interpreting the effects. We vary the data generation setting with respect to sample size, true number of data clusters and effect of the time covariate and generate 200 data sets for each data setting.

A joint model for all outcome variables is estimated assuming that the true number of data clusters is known. This is achieved by setting $G_{\mathsf{max}} = G$ and using $a_e = 4$ and $b_e = 1$ for the hyperparameters of the prior on the component weights to avoid sparse cluster solutions. Using this specification ensures that we estimate exactly G data clusters for each of the 200 simulated data sets. Posterior medians of the estimated group-specific intercepts are used to match the labelling of the estimates for the simulated data sets to the labels of the clusters used in data generation.

Figure 6.4 shows the results obtained for the slope estimates of the numeric and the binary outcome. The numeric outcome are much more informative, which results in remarkably thinner credible intervals. On the other hand, the binary outcome variable corresponds to the least informative outcome type and, thus, these results demonstrate that accurate estimation is achieved even under the most challenging conditions when the sample size is sufficiently large. Estimating a model with a common slope for all clusters leads to the correct estimation of the value 0 (in case no effect of time is present) or 2 (in case the clusters share the same slope term) for a sample size n of 250 or higher for $G = 2$ and 500 or higher for $G = 3$. However, an average effect is estimated when clusters indeed have a different slope. On the other hand, when estimating the model with different slopes across clusters, the group-specific estimates also coincide with the true common value (0 when no effect and 2 in the parallel lines), although a small shrinkage towards zero is visible for a low sample size n. Such a shrinkage behaviour can also be discerned in case the data generating process has group-specific slopes. However, this effect vanishes with increasing sample size and excellent results are obtained for $n = 1000$.



(a) The slope terms for the numeric outcome.

(b) The slope terms for the binary outcome.

Figure 6.4: Medians, 2.5 and 97.5% quantiles of estimated posterior medians of the slope term for the Num and Bin outcome variable across 200 simulated data sets. Model estimation is performed assuming that the number of data clusters G is known. Different settings are considered for the effect of the time covariate x for data generation (rows) and model specification (columns) and $\xi = \frac{1}{3}$ is used for data generation. The dashed lines indicate the true values. These are grey in case the effects are identical across clusters and in colour otherwise.

7. Implementation and numerical approximation methods

Theoretical proposition of a statistical model in itself is not enough for a practical application. The software implementation is an integral part of the development of a new methodology. Therefore, we dedicate this chapter to briefly introduce some of the key aspects of our implementation within the free statistical software R (R Core Team, 2022). Although there is a lot of potentially interesting details regarding the implementation, we restrict ourselves here to those interesting even from the theoretical point of view.

Since the focus of this chapter is more practical, we also deal with several technical issues that were intentionally postponed. In Chapter 2 we built the model around latent quantities that have to be integrated out to obtain the marginal pdfs (2.17) for the modelled outcomes. These are required for the posterior classification probabilities (3.4) viewed as a parametric function of model parameters $\boldsymbol{\theta}$. However, the necessary integrations have not been performed yet. The reason is simple, the exact evaluation of the integrals is practically impossible and only approximative methods have to be employed. We present algorithmic solutions to these problems including some of the well-known and broadly used algorithms (methods) that have been mentioned throughout the thesis for their diverse applications.

7.1 Implementation

First drafts of our implementation for the *threshold concept* were sketched solely within the free statistical software R (R Core Team, 2022) with the use of basic functions available. We allowed the user to choose which unknown model parameters should be group-specific and which should be common to all clusters. This required derivation of the full-conditional distributions under all possible combinations by slightly adjusting the formulas in Section 5.1. The implementation was in the end very flexible, compatible with `coda` package (Plummer et al., 2006) to obtain an MCMC summary and provided with several helpful plotting tools, see Figure 4.2 for an exemplary monitoring output.

However, the sampling algorithm was painfully slow for datasets of large sample size n. Hence, we tried to utilize some traditional software for MCMC, such as BUGS (Lunn et al., 2000), JAGS (Plummer, 2003; Denwood, 2016) or RStan (Stan Development Team, 2020). However, despite their user-friendly environment and a successful implementation for datasets of medium size, the waiting time for the EU-SILC data analysis was still unbearably slow (even a day of non-stop computation was not enough). Moreover, these samplers may use some advanced sampling techniques compared to the ones proposed. For example, we suspect them from stacking all the random effects (across all units) into one vector and then computing derivatives wrt this extremely long vector instead of updating them independently for all units.

Since we wanted to be sure how exactly the sampler proposes new values, we returned back to our implementation in R and tried to identify the bottle neck.

We noticed that the for cycle for sampling new draws from the full-conditional distributions (see Algorithms 1 and 2) could be substantially improved by replacing it with analogous implementation using the C language. Therefore, we kept the main structure of already implemented solution in R, which includes the data preparation phase and initialization steps, and then replaced the for cycle by calling a C function which would do exactly the same steps but much faster. The tricky part was the transfer of all needed data, parameters, dimensions to and then back from the C environment since their size changes with many specifications the analyst may require such as different formulas for the predictors or decision about cluster-specificity of the model parameters.

In the end, such a combination of R and C indeed proved to be about 80× more efficient than our original pure R solution. Pleased by this success, we implemented the calculation of classification probabilities and the model deviance in a similar fashion; data preparation and processing in R but the crucial calculations with the use of C functions.

For those interested in applications, the implementations for both the *threshold concept* model and the *GLMM-based* model are provided via GitHub at `https://github.com/vavrajan/` together with a tutorial on how to use them.

During the implementation of the second model we were able to improve the functionality in many aspects. We incorporated the sparse finite mixture approach which has to account for empty clusters and final post-processing of the data (see Section 6.3 for details). Moreover, the missing outcome values could be considered as additional model parameters to be estimated, which prevents the waste of incomplete data. But the most significant improvement was wrt the construction of the predictor.

Predictor construction for the GLMM-based model

The way one specifies the formula for the fixed and the random effects is rather primitive for the initial *threshold concept* model. The analyst has to supply the dataset of covariates and the set of column names to form the predictor, which means the regression matrix has to be created manually ahead. For the implementation of the *GLMM-based* model we made the interface more user friendly by utilizing the traditional specification in R by `formula`.

However, during the initial analysis of the EU-SILC dataset we observed that many regressors potentially effect the outcomes, but only few of them were relevant with respect to the clustering. Hence, there was an idea to separate the fixed part of the predictor η^{F} into a part common to all clusters η^{F} (still denoted by the same symbol) and a group-specific part η^{G}. With the presence of a count outcome and a possibility for having an offset η^{O} (by default $\eta^{\mathsf{O}} = 0$) the final structure of the predictor is supposed to be of the form:

$$\begin{aligned}\eta_{i,j}^{r,(g)} = \eta_{i,j}^{\mathsf{O},r} + \eta_{i,j}^{\mathsf{F},r} + \eta_{i,j}^{\mathsf{G},r,(g)} + \eta_{i,j}^{\mathsf{R},r} = \\ = \underbrace{\log o_{i,j}^{r}}_{\text{offset}} + \underbrace{\left(x_{i,j}^{\mathsf{F},r}\right)^{\top} \boldsymbol{\beta}_{r}^{\mathsf{F}}}_{\text{common fixed effects}} + \underbrace{\left(x_{i,j}^{\mathsf{G},r}\right)^{\top} \boldsymbol{\beta}_{r}^{(g)}}_{\text{group-specific effects}} + \underbrace{\left(z_{i,j}^{r}\right)^{\top} \boldsymbol{b}_{i}^{r}}_{\text{random effects}},\end{aligned} \quad (7.1)$$

where the covariates $x_{i,j}^{\mathsf{F},r}$ and $x_{i,j}^{\mathsf{G},r}$ are mutually exclusive (in case of collision the group-specific part has a preference). Effects $\boldsymbol{\beta}_{r}^{\mathsf{F}} \in \mathbb{R}^{d_r^{\mathsf{F}}}$ are common to all units $i = 1, \dots, n$, while $\boldsymbol{\beta}_r^{(g)} \in \mathbb{R}^{d_r^{\mathsf{G}}}$ are specific only for units within cluster g, i.e. $i \in \mathcal{N}_g(\boldsymbol{U})$. Note that these effects are allowed to be empty (either $d_r^{\mathsf{F}} = 0$ or $d_r^{\mathsf{G}} = 0$). We also have to adjust the overall notation for the effects: $\boldsymbol{\beta}^{\mathsf{F}} = \left\{\boldsymbol{\beta}_r^{\mathsf{F}}, r \in \mathcal{R}\right\}$, $\boldsymbol{\beta}^{(g)} = \left\{\boldsymbol{\beta}_r^{(g)}, r \in \mathcal{R}\right\}$, $\boldsymbol{\beta}_r = \left\{\boldsymbol{\beta}_r^{(g)}, g = 1, \dots, G\right\}$ and $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}_r^{(g)}, r \in \mathcal{R}, g = 1, \dots, G\right\}$. The notation for general categorical outcome would require an additional subscript k for the corresponding outcome level, which is going to be ignored from now on to focus on the main goal.

From the theoretical point of view, we have to derive full-conditional distributions for $\boldsymbol{\beta}_r^{\mathsf{F}}$ and $\boldsymbol{\beta}_r^{(g)}$ separately. In Section 5.1.6, we derived the full-conditional distribution when the underlying distribution of the outcomes is assumed to be normal, $r \in \mathcal{R}^{\mathsf{Num}}$. In this case the prior for the effects has to be adjusted as well depending on the group-specificity of the precision to not to violate the hierarchical structure of the model:

- $\tau_r^{(g)}$ is group-specific, then only $p\left(\boldsymbol{\beta}_r^{(g)} \,\middle|\, \tau_r^{(g)}\right) p\left(\tau_r^{(g)}\right)$ is assumed while τ-free prior $p\left(\boldsymbol{\beta}_r^{\mathsf{F}} \,\middle|\, \boldsymbol{\beta}_{0,r}^{\mathsf{F}}, \mathbb{D}_{r,\mathsf{F}}\right)$ was used similarly as for non-numeric outcomes,
- τ_r is common to all clusters, then priors for both $\boldsymbol{\beta}_r^{\mathsf{F}}$ and $\boldsymbol{\beta}_r^{(g)}$ can be tied with τ_r,

see Section 4.2 to refresh the details of the prior setting.

Another key moment for derivation of the full-conditional distributions is when the shifted outcomes $\widetilde{Y}_{i,j}^r$ are created. When $\boldsymbol{\beta}_r^{\mathsf{F}}$ is the primary target then

$$\widetilde{Y}_{i,j}^r := Y_{i,j}^r - \eta_{i,j}^{\mathsf{O},r} - \eta_{i,j}^{\mathsf{G},r,(U_i)} - \eta_{i,j}^{\mathsf{R},r}.$$

where the group-specific part of the predictor is defined by the cluster $g = U_i$ unit i currently belongs to. Similarly, the precision $\tau_r^{(U_i)}$ (or common τ_r) is used. Its full-conditional distribution is then derived from all units $i = 1, \dots, n$. On the other hand, when $\boldsymbol{\beta}_r^{(g)}$ is of the interest then

$$\widetilde{Y}_{i,j}^r := Y_{i,j}^r - \eta_{i,j}^{\mathsf{O},r} - \eta_{i,j}^{\mathsf{F},r} - \eta_{i,j}^{\mathsf{R},r}$$

for any group g, however, only units i currently within the cluster g contribute to the full-conditional distribution of $\boldsymbol{\beta}_r^{(g)}$ the same way as in Section 5.1.6.

In case of non-numeric outcomes $r \in \mathcal{R} \setminus \mathcal{R}^{\mathsf{Num}}$, the derivatives of the log-pdfs $\ell(\bullet|\cdots)$ of the full-conditional distributions (see Section 6.2.2) have to be done also separately for $\boldsymbol{\beta}_r^{\mathsf{F}}$ and $\boldsymbol{\beta}_r^{(g)}$, but will remain almost the same. The key difference here are the derivatives of the predictor wrt the effects:

$$\frac{\partial \eta_{i,j}^{r,(g)}}{\partial \boldsymbol{\beta}_r^{\mathsf{F}}} = x_{i,j}^{\mathsf{F},r}, \quad \frac{\partial \eta_{i,j}^{r,(g)}}{\partial \boldsymbol{\beta}_r^{(g')}} = \mathbb{1}_{(g=g')} x_{i,j}^{\mathsf{G},r} \quad \text{and} \quad \frac{\partial \eta_{i,j}^{r,(g)}}{\partial \boldsymbol{b}_i^r} = z_{i,j}^r.$$

From the practical point of view, this division of the predictor into four disjoint parts is an opportunity for its efficient evaluation provided that sufficient memory capacity is available. Each $\eta_{i,j}^{r,(g)}$ can be represented by four memory slots corresponding to $\eta_{i,j}^{\mathsf{O},r}, \eta_{i,j}^{\mathsf{F},r}, \eta_{i,j}^{\mathsf{G},r,(g)}$ and $\eta_{i,j}^{\mathsf{R},r}$, where the offset and linear combination of the covariates with the last known values of $\boldsymbol{\beta}_r^{\mathsf{F}}$, $\boldsymbol{\beta}_r^{(g)}$ and $\boldsymbol{b}_i^r$ are stored. When the overall value of the predictor $\eta_{i,j}^{r,(g)}$ is needed, it suffices to add up these four stored numbers. When evaluating $\ell(\bullet|\cdots)$ and its derivatives, only one of the memory slots has to be updated (e.g. within the iterative process of Newton–Raphson method, see Algorithm 6) while the other ones remain intact, beauty of which comes from the full-conditionality. Therefore, a lot of computational time that would have been otherwise spent on evaluation of the predictor is saved.

Nevertheless, each time a new value of fixed, group-specific or random effects is drawn the corresponding memory slots have to be updated immediately afterwards. Moreover, there is practically only one slot for the group-specific part $\eta_{i,j}^{\mathsf{G},r,(g)}$ (not G of them as the notation suggests) which has to be interpreted as the group-specific part of the predictor for observation j of outcome r of unit i in its current cluster $U_i = g$. Meaning that this group-specific slot has to be updated even after the allocation indicators U_i are sampled; at least those units i which change their cluster allocation: $U_i^m \neq U_i^{m-1}$.

To sum it up, Algorithm 2 should be extended by

- initialization of the memory slots of the predictor,
- additional sampling step of the effects common to all clusters $\boldsymbol{\beta}_r^{\mathsf{F}}$ and
- updates of the predictor after sampling each $\boldsymbol{\beta}_r^{\mathsf{F}}$, $\boldsymbol{\beta}_r^{(g)}$, $\boldsymbol{b}_i$ and U_i.

PBC910 example

We illustrate the capabilities of our implementation on the PBC910 example used in Sections 2.4.1, 2.4.2, 3.3.1 and 3.3.2. The fixed part of the predictor consisted of the spline parametrization of time since the entry and the interaction between age and gender (2.22). There was also a random intercept term specific to each patient.

In Section 2.4, only one cluster $G = 1$ is assumed, hence, it does not matter whether the effects are considered common to all clusters or group-specific. However, imagine a situation $G > 1$ with the requirement that the effects would still be the same for all patients. In R notation one would have to specify:

$$\begin{aligned} &\texttt{fixed} \sim \texttt{age} * \texttt{sex} + \texttt{bs}(\texttt{time}, \texttt{knots}, \ldots), &&\texttt{random} \sim 1,\\ &\texttt{group} \sim -1, &&\texttt{offset} = 0. \end{aligned}$$

On the other hand, in Section 3.3 all the fixed effects were considered to be group-specific, which was achieved by

$$\begin{aligned} &\texttt{fixed} \sim -1, &&\texttt{random} \sim 1,\\ &\texttt{group} \sim \texttt{age} * \texttt{sex} + \texttt{bs}(\texttt{time}, \texttt{knots}, \ldots), &&\texttt{offset} = 0. \end{aligned}$$

We could even leave the `fixed` formula as it was since when the two terms in `fixed` and `group` coincide, the group-specificity has preference by default.

Suppose we would be interested only in the differences caused by the evolution in time and not the effects of other covariates. That is,

$$\begin{aligned} \eta = \underbrace{0}_{\text{offset}} + \underbrace{\beta_A A + \beta_M M + \beta_{A:M} AM}_{\text{common fixed part}} + \\ + \underbrace{\beta_0^{(g)} + \beta_1^{(g)} S_1 + \beta_2^{(g)} S_2 + \beta_3^{(g)} S_3}_{\text{group-specific part}} + \underbrace{\quad b_i \quad}_{\text{random intercept}}. \end{aligned}$$

Table 7.1: `PBC910` dataset, $G = 2$. Posterior medians of the *GLMM-based* model parameters including 95% equal-tailed credible intervals. Effects of age and sex are common to both clusters.

Parameter	g	Numeric outcome		Count outcome		Binary outcome		Ordinal outcome	
		Log(bilirubin)		Platelet count		Hepatomegaly		Edema	
β_0	1	1.27	(0.55; 1.77)	5.57	(5.46; 5.72)	0.55	(−2.11; 2.52)	-	
	2	0.82	(0.10; 1.34)	5.56	(5.45; 5.69)	−0.45	(−3.04; 1.51)	-	
β_A	1=2	−0.14	(−0.24; 0.01)	0.00	(−0.03; 0.02)	−0.08	(−0.46; 0.41)	0.71	(0.22; 1.23)
β_M	1=2	0.33	(−0.99; 1.62)	0.29	(−0.40; 0.92)	−0.25	(−2.10; 1.61)	−1.05	(−2.45; 0.35)
$\beta_{A:M}$	1=2	0.03	(−0.20; 0.27)	−0.07	(−0.19; 0.06)	0.32	(−0.10; 0.76)	0.01	(−1.13; 1.15)
β_1	1	−0.10	(−0.28; 0.07)	−0.33	(−0.36; −0.29)	0.44	(−1.05; 1.87)	−0.58	(−1.97; 0.85)
	2	−0.13	(−0.27; 0.02)	0.00	(−0.03; 0.03)	−0.51	(−1.68; 0.64)	−0.60	(−2.12; 0.93)
β_2	1	0.02	(−0.24; 0.30)	−0.27	(−0.33; −0.21)	−0.99	(−3.35; 1.47)	1.30	(−0.69; 3.38)
	2	0.34	(0.11; 0.56)	0.12	(0.09; 0.16)	1.30	(−0.44; 3.11)	2.12	(−0.30; 4.58)
β_3	1	0.42	(0.08; 0.76)	−0.66	(−0.73; −0.58)	3.12	(−0.32; 6.51)	1.82	(−0.56; 4.26)
	2	0.07	(−0.20; 0.35)	0.07	(0.02; 0.11)	−1.26	(−3.66; 0.95)	−0.44	(−3.88; 2.68)
$\sigma = \tau^{-\frac{1}{2}}$	1	0.39	(0.36; 0.43)	-		-		-	
	2	0.37	(0.35; 0.40)	-		-		-	
c_0	1	-		-		-		2.89	(1.95; 3.83)
	2	-		-		-		3.54	(2.54; 4.69)
c_1	1	-		-		-		6.39	(5.09; 7.78)
	2	-		-		-		7.71	(6.26; 9.52)

Then, we would have to set up the formulae in the following way:

$$\texttt{fixed} \sim \texttt{age} * \texttt{sex}, \qquad \texttt{random} \sim 1,$$
$$\texttt{group} \sim \texttt{bs}(\texttt{time}, \texttt{knots}, \ldots), \qquad \texttt{offset} = 0.$$

We will split the fixed effects similarly in the analysis of the EU-SILC dataset in Chapter 8.

We have reestimated this model under this setting and created Table 7.1 summarizing the posterior of the model parameters. The resulting estimates of group-specific parameters are somewhat analogous to those in Table 3.2, while the estimates of effects common to both clusters remind the estimates from Table 2.2. Moreover, we created Figure 7.1 with analogous plots as in Figure 3.3. We can clearly see that the evolution in time is different for both clusters, however, the effects of age and sex are comparable between these two clusters.

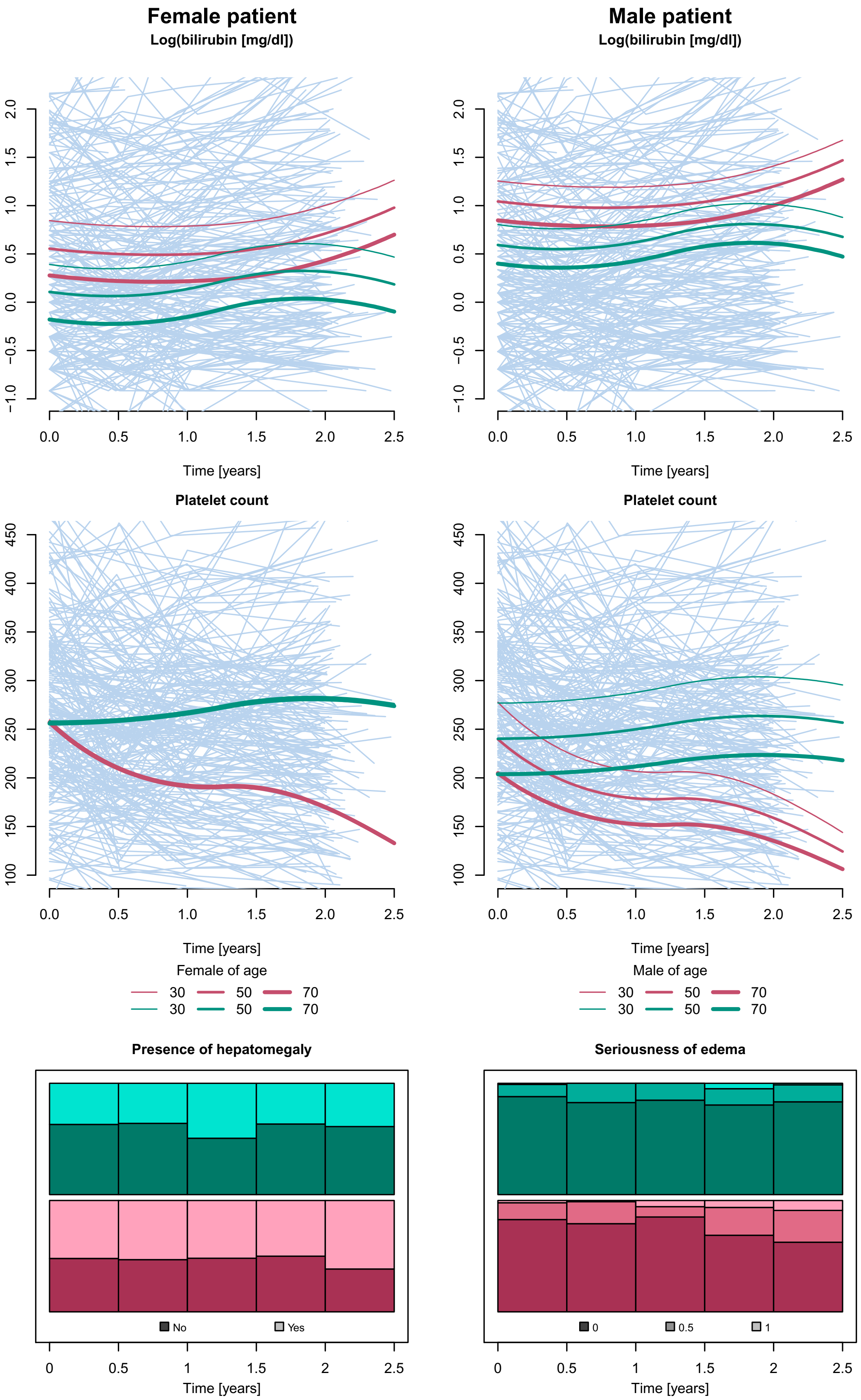


Figure 7.1: PBC910 dataset. *GLMM-based* model, $G = 2$. Estimated group-specific (red ($g = 1$) and turquoise ($g = 2$)) spline curves by posterior median. Effects of age and sex are assumed to be common to both clusters. Proportions of categorical outcomes with respect to time separately in each cluster.

7.2 Evaluation of the pdf in the threshold concept model

In Section 2.4.1 we outlined the *threshold concept* model given by the pdf (2.18). The mixture of these models (3.5) was created later in Section 3.3.1, where we noted that if one wishes to evaluate the posterior allocation probability $u_{i,g}(\boldsymbol{\theta})$ given by (3.6), the integrals in (2.18) have to be evaluated for all G clusters. Moreover, in Section 4.3 we introduced deviance of the model, which requires evaluation of the same quantities for all units i within the training dataset.

To be precise, Section 4.3 also introduces randomized hyperparameters $\mathcal{H}$ (only $\mathbb{Q}$ for this model) for the prior of the unknown parameters $\boldsymbol{\theta}$, which altogether creates the posterior (4.19). Neither outcomes nor latent quantities depend on these randomized hyperparameters, only on $\boldsymbol{\theta}$. Integration of $\mathcal{H}$ from $p(\boldsymbol{\theta}, \mathcal{H}) = p(\boldsymbol{\theta}|\mathcal{H})p(\mathcal{H})$ leads to marginalized prior $p(\boldsymbol{\theta})$ without having any effect on the rest of the model. Therefore, we can neglect this part of the model and focus on the integration of other important latent quantities to obtain the marginal pdf for observed outcomes.

Within this section, we provide a guide on how to perform the integration in (2.18) with respect to auxiliary latent variables (random effects $\boldsymbol{b}_i$ and the latent numeric outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$) in order to evaluate the classification probabilities $u_{i,g}(\boldsymbol{\theta})$ and deviance $D^G(\boldsymbol{\theta}; \mathbb{Y}, \mathcal{C})$. To recall (2.18), which consists of (2.19), (2.20) and (2.21), we reshape it into a pdf for the cluster g:

$$p\left(\mathbb{Y}_i \,\middle|\, U_i = g, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \gamma, \boldsymbol{\Sigma}^{(g)}; \mathcal{C}_i\right) = \int \underbrace{p\left(\mathbb{Y}_i^{\mathsf{OB}} \,\middle|\, \mathbb{Y}_i^{\star,\mathsf{OB}}, \gamma\right)}_{\text{threshold concept (2.19)}} \cdot$$
$$\cdot \left[\int \underbrace{p\left(\mathbb{Y}_i^{\mathsf{N}} \,\middle|\, \boldsymbol{b}_i^{\mathsf{N}}, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}; \mathcal{C}_i\right) p\left(\mathbb{Y}_i^{\star,\mathsf{OB}} \,\middle|\, \boldsymbol{b}_i^{\mathsf{OB}}, \boldsymbol{\beta}^{(g)}; \mathcal{C}_i\right)}_{\text{multivariate LME (2.20)}} \cdot \underbrace{p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\Sigma}^{(g)}\right)}_{(2.21)} \mathrm{d}\boldsymbol{b}_i\right] \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}, \tag{7.2}$$

where we inserted the parameters specific for cluster g.

Integration with respect to random effects $\boldsymbol{b}_i$

Let us first integrate the random effects $\boldsymbol{b}_i$ out of (7.2) to obtain the marginal distribution of numeric variables. In this case, we will avoid integration by realization that under the normality assumption of both numeric outcomes and random effects the unconditional distribution of the outcomes is also normal.

Let us gather the observed numeric outcomes $\mathbb{Y}_i^{\mathsf{N}}$ and the latent numeric outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$ into one vector $\boldsymbol{Y}_i$ of length $d = n_i|\mathcal{R}|$. Then, $\boldsymbol{Y}_i$ given a vector of all random effects $\boldsymbol{b}_i$ follows by our LME assumption (2.1) and (2.5) multivariate normal distribution:

$$\boldsymbol{Y}_i \,\middle|\, \boldsymbol{b}_i,\, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{\Sigma}^{(g)};\, \mathcal{C}_i \sim \mathsf{N}_d\left(\mathbb{X}_i\boldsymbol{\beta}^{(g)} + \mathbb{Z}_i\boldsymbol{b}_i,\, \mathbb{T}_i^{(g)}\right),$$

where $\mathbb{X}_i$ and $\mathbb{Z}_i$ are block diagonal matrices composed of model matrices of fixed effects $\mathbb{X}_i^r$ and of random effects $\mathbb{Z}_i^r$, respectively. The covariance matrix $\mathbb{T}_i^{(g)}$

is diagonal due to the independence assumption and contains the corresponding parameters of the residual variability, that is, $\left(\tau_r^{(g)}\right)^{-1}$ for $r \in \mathcal{R}^{\mathsf{Num}}$ and 1 otherwise.

Similarly as in the introduction to Chapter 2, we compute the marginal mean and variance matrix of $\boldsymbol{Y}_i$ unconditioned by the random effects $\boldsymbol{b}_i$:

$$\begin{aligned}
\mathsf{E}\left[\boldsymbol{Y}_i \,\middle|\, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{\Sigma}^{(g)};\, \mathcal{C}_i\right] &= \mathsf{E}\Big(\mathsf{E}\Big[\boldsymbol{Y}_i\Big|\boldsymbol{b}_i, \ldots\Big]\Big) = \mathbb{X}_i\boldsymbol{\beta}^{(g)} + \mathbb{Z}_i\mathbf{0} = \mathbb{X}_i\boldsymbol{\beta}^{(g)} =: \boldsymbol{\mu}_i^{(g)}, \\
\mathsf{var}\left[\boldsymbol{Y}_i \,\middle|\, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{\Sigma}^{(g)};\, \mathcal{C}_i\right] &= \mathsf{E}\Big(\mathsf{var}\Big[\boldsymbol{Y}_i\Big|\boldsymbol{b}_i, \ldots\Big]\Big) + \mathsf{var}\Big(\mathsf{E}\Big[\boldsymbol{Y}_i\Big|\boldsymbol{b}_i, \ldots\Big]\Big) \\
&= \mathbb{T}_i^{(g)} + \mathbb{Z}_i^\top \boldsymbol{\Sigma}^{(g)} \mathbb{Z}_i =: \mathbb{V}_i^{(g)}.
\end{aligned}$$

Due to conjugacy of normal distributions for $\boldsymbol{Y}_i|\boldsymbol{b}_i$ and $\boldsymbol{b}_i$, the marginal distribution of $\boldsymbol{Y}_i$ has to preserve the normality:

$$\boldsymbol{Y}_i \,\Big|\, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{\Sigma}^{(g)};\, \mathcal{C}_i \sim \mathsf{N}_d\left(\boldsymbol{\mu}_i^{(g)},\, \mathbb{V}_i^{(g)}\right). \tag{7.3}$$

This distribution has a general covariance structure, which is a result of modelling outcomes jointly through random effects (Section 2.4) to capture dependencies among the longitudinal outcomes. In the following, we will need to divide this distribution into parts corresponding to numeric and categorical (ordinal and binary) outcomes, hence, the notation:

$$\boldsymbol{\mu}_i^{(g)} = \begin{pmatrix} \boldsymbol{\mu}_{i,\mathsf{N}}^{(g)} \\ \boldsymbol{\mu}_{i,\mathsf{OB}}^{(g)} \end{pmatrix} \quad \text{and} \quad \mathbb{V}_i^{(g)} = \begin{pmatrix} \mathbb{V}_{i,\mathsf{N}}^{(g)} & \mathbb{V}_{i,\mathsf{NOB}}^{(g)} \\ \mathbb{V}_{i,\mathsf{OBN}}^{(g)} & \mathbb{V}_{i,\mathsf{OB}}^{(g)} \end{pmatrix}.$$

Integration with respect to latent numeric outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$

It remains to perform the following integration:

$$\int p\left(\mathbb{Y}_i^{\mathsf{OB}} \,\Big|\, \mathbb{Y}_i^{\star,\mathsf{OB}},\, \boldsymbol{\gamma}\right) \cdot p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \,\Big|\, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{\Sigma}^{(g)};\, \mathcal{C}_i\right) \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}},$$

which is, in fact, an integration of a multivariate normal density within the bounds given by the thresholds $\boldsymbol{\gamma}$ and the observed ordinal and binary outcomes $\mathbb{Y}_i^{\mathsf{OB}}$. First, we separate marginal distribution of numeric outcomes $\mathbb{Y}_i^{\mathsf{N}}$ since it can avoid the integration. However, now after integration of $\boldsymbol{b}_i$ the outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$ and $\mathbb{Y}_i^{\mathsf{N}}$ are dependent. Hence, the conditional normal distribution of latent numeric outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$ given $\mathbb{Y}_i^{\mathsf{N}}$ still awaits the integration:

$$\underbrace{\varphi\left(\mathbb{Y}_i^{\mathsf{N}};\, \boldsymbol{\mu}_{i,\mathsf{N}}^{(g)}, \mathbb{V}_{i,\mathsf{N}}^{(g)}\right)}_{\text{pdf of vectorized } \mathbb{Y}_i^{\mathsf{N}}} \cdot \int \underbrace{p\left(\mathbb{Y}_i^{\mathsf{OB}} \,\Big|\, \mathbb{Y}_i^{\star,\mathsf{OB}},\, \boldsymbol{\gamma}\right)}_{\text{thresholding (2.19)}} \cdot \underbrace{\varphi\left(\mathbb{Y}_i^{\star,\mathsf{OB}};\, \boldsymbol{\mu}_{i,\mathsf{OB|N}}^{(g)}, \mathbb{V}_{i,\mathsf{OB|N}}^{(g)}\right)}_{\text{pdf of } \mathbb{Y}_i^{\star,\mathsf{OB}} \big| \mathbb{Y}_i^{\mathsf{N}}} \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}},$$

where $\boldsymbol{\mu}_{i,\mathsf{OB|N}}^{(g)}$ and $\mathbb{V}_{i,\mathsf{OB|N}}^{(g)}$ are the conditional mean and the covariance matrix of $\mathbb{Y}_i^{\star,\mathsf{OB}} \,\Big|\, \mathbb{Y}_i^{\mathsf{N}},\, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{\Sigma}^{(g)};\, \mathcal{C}_i$ and are given by the standard formulae:

$$\begin{aligned}
\boldsymbol{\mu}_{i,\mathsf{OB|N}}^{(g)} &= \boldsymbol{\mu}_{i,\mathsf{OB}}^{(g)} + \mathbb{V}_{i,\mathsf{OBN}}^{(g)} \left(\mathbb{V}_{i,\mathsf{N}}^{(g)}\right)^{-1} \left(\mathbb{Y}_i^{\mathsf{N}} - \boldsymbol{\mu}_{i,\mathsf{N}}^{(g)}\right), \\
\mathbb{V}_{i,\mathsf{OB|N}}^{(g)} &= \mathbb{V}_{i,\mathsf{OB}}^{(g)} - \mathbb{V}_{i,\mathsf{OBN}}^{(g)} \left(\mathbb{V}_{i,\mathsf{N}}^{(g)}\right)^{-1} \mathbb{V}_{i,\mathsf{NOB}}^{(g)}.
\end{aligned}$$

It remains to integrate the product of two functions, the first of which only declares lower and upper integration bounds while the second is the probability density function of a multivariate normal distribution with the mean $\boldsymbol{\eta}^{(g)}_{i,\mathsf{OB|N}}$ and the covariance matrix $\mathbb{V}^{(g)}_{i,\mathsf{OB|N}}$. For each individual categorical outcome $r \in \mathcal{R}^{\mathsf{OB}}$ and observation $j \in \{1, \dots, n_i\}$ the value $Y^r_{i,j} = k$ determines an interval given by the corresponding pair of $\boldsymbol{\gamma}$ parameters, see (2.4):

$$Y^r_{i,j} = k \quad \Longrightarrow \quad Y^{\star,r}_{i,j} \in \left(\gamma^r_{k-1},\ \gamma^r_k\right] =: \left(e^r_{i,j}, f^r_{i,j}\right].$$

If we denote the resulting Cartesian product of these intervals as $\square\left(\boldsymbol{\gamma},\ \mathbb{Y}^{\mathsf{OB}}_i\right) = (\boldsymbol{e}_i, \boldsymbol{f}_i] \subset \mathbb{R}^{d^{\mathsf{OB}}}$ then the remaining integral can be written in the form

$$I_g\left(\mathbb{Y}^{\mathsf{OB}}_i\right) := \int\limits_{\square\left(\boldsymbol{\gamma},\, \mathbb{Y}^{\mathsf{OB}}_i\right)} p\left(\mathbb{Y}^{\star,\mathsf{OB}}_i \,\middle|\, \mathbb{Y}^{\mathsf{N}}_i, \boldsymbol{\beta}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{\Sigma}^{(g)}; \mathcal{C}_i\right) \mathrm{d}\mathbb{Y}^{\star,\mathsf{OB}}_i =$$

$$= \int\limits_{\boldsymbol{e}_i}^{\boldsymbol{f}_i} \varphi\left(\boldsymbol{y};\ \boldsymbol{\mu}^{(g)}_{i,\mathsf{OB|N}}, \mathbb{V}^{(g)}_{i,\mathsf{OB|N}}\right) \mathrm{d}\boldsymbol{y}. \qquad (7.4)$$

Finally, after the integrals I_g for all $g = 1, \dots, G$ are computed, the classification probabilities can be calculated proportionally:

$$u_{i,g}\left(\boldsymbol{\theta}\right) = \frac{w_g \cdot \varphi\left(\mathbb{Y}^{\mathsf{N}}_i;\ \boldsymbol{\mu}^{(g)}_{i,\mathsf{N}}, \mathbb{V}^{(g)}_{i,\mathsf{N}}\right) \cdot I_g\left(\mathbb{Y}^{\mathsf{OB}}_i\right)}{\sum\limits_{g'=1}^{G} w_{g'} \cdot \varphi\left(\mathbb{Y}^{\mathsf{N}}_i;\ \boldsymbol{\mu}^{(g')}_{i,\mathsf{N}}, \mathbb{V}^{(g')}_{i,\mathsf{N}}\right) \cdot I_{g'}\left(\mathbb{Y}^{\mathsf{OB}}_i\right)}. \qquad (7.5)$$

Similarly as for full-conditional distributions, it is computationally more stable to evaluate the numerators on a log-scale first. Then shift them all by the same suitable constant, which will result in exponentials of reasonable scale to be compared in the fraction. However, this is not applicable for the contribution of unit i to the deviance (4.18), which is given directly by the denominator of (7.5). Any constants lost due to proportionality should be retrieved. If the use of deviance is to compare two different models, constants that would be common to both models could be skipped.

The d^{OB}-dimensional integrals (7.4) needed in (7.5) could be directly evaluated through $2^{d^{\mathsf{OB}}}$ values of cumulative distribution function of multivariate normal distribution. With $d^{\mathsf{OB}} = n_i|\mathcal{R}^{\mathsf{OB}}|$ large, the evaluation becomes very expensive, hence, a different approximative approach is needed. For that reason, we adopted an effective algorithm presented by Genz (1992) which is also based on the MCMC sampling, see Algorithm 4. To obtain I_g, the procedure MultNormProb would be called for $d = d^{\mathsf{OB}}$, $\mathbb{V} = \mathbb{V}^{(g)}_{i,\mathsf{OB|N}}$, $\boldsymbol{e} = \boldsymbol{e}_i - \boldsymbol{\mu}^{(g)}_{i,\mathsf{OB|N}}$ and $\boldsymbol{f} = \boldsymbol{f}_i - \boldsymbol{\mu}^{(g)}_{i,\mathsf{OB|N}}$, where the underlying distribution is centred to fit the assumption of the procedure. The implemented function `pmvnorm` from the R package `mvtnorm` (Genz et al., 2020) is used in our applications.

Since the approximation of such an integral is needed G-times for each generated state of the Gibbs sampling, the overall procedure is still considerably time-consuming.

Algorithm 4 Approximation of multivariate normal probabilities by Genz (1992)

```
Used functions:
$\textsc{CholeskySolve}(A, \boldsymbol{b})$ ▷ returning $\mathbb{C}$ from $A = \mathbb{C}\mathbb{C}^\top$ and solution to $\mathbb{C}\boldsymbol{x} \overset{!}{=} \boldsymbol{b}$
$\Phi(x)$                ▷ cumulative distribution function of $\mathsf{N}(0,\, 1)$
$\Phi^{-1}(p)$           ▷ quantile function of $\mathsf{N}(0,\, 1)$

procedure $\textsc{MultNormProb}(\mathbb{V}, \boldsymbol{e}, \boldsymbol{f}, \epsilon, \alpha, N_{\mathsf{max}})$
    ▷ $\epsilon$ tolerance for error, $\alpha$ Monte Carlo confidence factor for standard error ($\alpha = 2.5$ for 99%), $N_{\mathsf{max}}$ total amount of iterations allowed for computation
    goal: $I = \mathsf{P}\left[\boldsymbol{e} < \mathsf{N}_d(\boldsymbol{0},\, \mathbb{V}) \le \boldsymbol{f}\right] = (2\pi)^{\frac{d}{2}}\, |\mathbb{V}|^{-\frac{1}{2}} \int_{\boldsymbol{e}}^{\boldsymbol{f}} \exp\left\{-\frac{1}{2}\boldsymbol{y}^\top \mathbb{V}^{-1}\boldsymbol{y}\right\}\, \mathrm{d}\boldsymbol{y}$.

    $\mathbb{C} \leftarrow \textsc{CholeskySolve}(\mathbb{V}, \cdot)$     ▷ only Cholesky decomposition required

    initialize: $\texttt{intsum} := 0,\ N := 0,\ \texttt{varsum} := 0;$
                $s_1 := \Phi(e_1/c_{1,1}),\ t_1 := \Phi(f_1/c_{1,1}),\ p_1 := t_1 - s_1;$
    repeat
        generate $w_1, \ldots, w_{d-1} \overset{\mathsf{iid}}{\sim} \mathsf{Unif}\,[0, 1]$;
        for $i$ in $2 : d$ do
            $y_{i-1} := \Phi^{-1}\left(s_{i-1} + w_{i-1}(t_{i-1} - d_{i-1})\right);$
            if $e_i = -\infty$ then
                $s_i := 0;$
            else
                $s_i := \Phi\left(\left(e_i - \sum_{j=1}^{i-1} c_{i,j} y_j\right) \big/ c_{i,i}\right);$
            end if
            if $f_i = \infty$ then
                $t_i := 1;$
            else
                $t_i := \Phi\left(\left(f_i - \sum_{j=1}^{i-1} c_{i,j} y_j\right) \big/ c_{i,i}\right);$
            end if
            $p_i := (t_i - s_i) p_{i-1};$
        end for
        $\texttt{intsum} := \texttt{intsum} + p_d;$
        $\texttt{varsum} := \texttt{varsum} + p_d^2;$
        $N \leftarrow N + 1;$
        $\texttt{error} := \alpha \sqrt{\left(\texttt{varsum}/N - (\texttt{intsum}/N)^2\right) \big/ N};$     ▷ MC error
    until $\texttt{error} < \epsilon$ or $N = N_{\mathsf{max}}$
    return $I = \texttt{intsum}/N$, $\texttt{error}$ and $N$.
end procedure
```

7.3 Classification probabilities for the GLMM-based model

The motivation for calculation of the classification probabilities for the *GLMM-based* model is analogous to the one in the previous section.

In Section 2.4.2, we proposed to combine different GLMM by a joint distribution of random effects $\boldsymbol{b}_i$, which resulted in the pdf (2.24). Later in Section 3.3.2, we created a mixture of these models (3.7) and gave the formula (3.8) for the posterior classification probability $u_{i,g}(\boldsymbol{\theta})$ involving the integral from (2.24). Adopting the Bayesian approach we added a prior distribution for $\boldsymbol{\theta}$ enriched by the randomized hyperparameters $\mathcal{H} = \{\mathbb{Q}, e_0\}$. Moreover, in Section 6.3 we even considered that some of the outcomes ($\mathbb{Y}_i^{\mathsf{mis}}$) may not be observed and are treated as additional latent model parameters by BDA.

These extensions force us to slow down and apply the Bayes' theorem once again to obtain the posterior classification probabilities

$$\begin{aligned}\mathsf{P}\left[U_i = g \,\middle|\, \mathbb{Y}_i^{\mathsf{obs}}, \boldsymbol{\theta}; \mathcal{C}_i\right] &\propto p\left(\mathbb{Y}_i^{\mathsf{obs}}, \boldsymbol{\theta} \,\middle|\, U_i = g; \mathcal{C}_i\right) \mathsf{P}\left[U_i = g\right]\\
&\propto \int\!\!\int\!\!\int w_g p\left(\mathbb{Y}_i^{\mathsf{obs}}, \mathbb{Y}_i^{\mathsf{mis}}, \boldsymbol{b}_i, \boldsymbol{\theta}, \mathcal{H} \,\middle|\, U_i = g; \mathcal{C}_i\right) \,\mathrm{d}\mathbb{Y}_i^{\mathsf{mis}} \,\mathrm{d}\boldsymbol{b}_i \,\mathrm{d}\mathcal{H}\\
&\propto w_g \int\!\!\int\!\!\int p\left(\mathbb{Y}_i^{\mathsf{obs}} \,\middle|\, U_i = g, \boldsymbol{b}_i, \boldsymbol{\theta}; \mathcal{C}_i\right) p\left(\mathbb{Y}_i^{\mathsf{mis}} \,\middle|\, U_i = g, \boldsymbol{b}_i, \boldsymbol{\theta}; \mathcal{C}_i\right) \cdot\\
&\qquad \cdot p\left(\boldsymbol{b}_i | U_i = g, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} | \mathcal{H}\right) p(\mathcal{H}) \,\mathrm{d}\mathbb{Y}_i^{\mathsf{mis}} \,\mathrm{d}\boldsymbol{b}_i \,\mathrm{d}\mathcal{H}.\end{aligned}$$

The missing outcome values appear only in one of the factors and are independent of $\mathbb{Y}_i^{\mathsf{obs}}$ given $\boldsymbol{b}_i$, hence, the integration wrt $\mathbb{Y}_i^{\mathsf{mis}}$ reduces to 1. Similarly as in the introduction to the previous section, we notice that integration wrt $\mathcal{H}$ only reduces the prior to $p(\boldsymbol{\theta})$, which does not depend on the fact that $U_i = g$. Therefore, it only becomes yet another multiplicative constant common to all clusters to be hidden within the proportionality sign $\propto$. Finally, we end up with slightly modified version of (3.8) which defines the clustering probabilities $u_{i,g}(\boldsymbol{\theta})$:

$$u_{i,g}(\boldsymbol{\theta}) \propto w_g \int \prod_{r\in\mathcal{R}} \prod_{j=1}^{n_i} \left[p_{\mathsf{t}(r)}\left(Y_{i,j}^r \middle| \boldsymbol{b}_i^r, \boldsymbol{\beta}_r^{(g)}, \tau_r^{(g)}, \boldsymbol{c}_r^{(g)}; \mathcal{C}_{i,j}\right)\right]^{I_{i,j}^r} \cdot p\left(\boldsymbol{b}_i \middle| \boldsymbol{\Sigma}^{(g)}\right) \,\mathrm{d}\boldsymbol{b}_i, \quad (7.6)$$

where only the observed outcome values contribute, which is denoted by indicators $I_{i,j}^r = \mathbb{1}_{(Y_{i,j}^r \text{ is observed})}$.

The integral in (7.6) could be fully factored in the fashion of (2.24) where parameters specific to group g would be used. More importantly, the proportionality (7.6) could be expressed in the following form:

$$u_{i,g}(\boldsymbol{\theta}) \propto w_g \left|\boldsymbol{\Sigma}^{(g)}\right|^{-\frac{1}{2}} \int \exp\left\{h_g(\boldsymbol{b}_i)\right\} \,\mathrm{d}\boldsymbol{b}_i, \quad (7.7)$$

where the function h_g is analogous to (6.11), although some summands, such as $\log\tau$, dependent on g are hidden in the additive constant as they were free of $\boldsymbol{b}_i$.

To be precise, the full expression of h_g in this case takes the form of

$$\begin{aligned}
h_g(\boldsymbol{b}_i) = -\frac{1}{2}\boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-(g)}\boldsymbol{b}_i + \sum_{r\in\mathcal{R}^{\mathsf{Num}}} \frac{\sum_{j=1}^{n_i} I^r_{i,j}}{2} \log \tau_r^{(g)} - &\\
- \sum_{r\in\mathcal{R}^{\mathsf{Num}}} \frac{\tau_r^{(g)}}{2} \sum_{j=1}^{n_i} I^r_{i,j}\left(Y^r_{i,j} - \eta^{r,(g)}_{i,j}\right)^2 + &\\
+ \sum_{r\in\mathcal{R}^{\mathsf{Poi}}} \sum_{j=1}^{n_i} I^r_{i,j}\left(Y^r_{i,j}\eta^{r,(g)}_{i,j} - \exp\left\{\eta^{r,(g)}_{i,j}\right\}\right) + &\\
+ \sum_{r\in\mathcal{R}^{\mathsf{Bin}}} \sum_{j=1}^{n_i} I^r_{i,j}\left[Y^r_{i,j}\eta^{r,(g)}_{i,j} - \log\left(1 + \exp\left\{\eta^{r,(g)}_{i,j}\right\}\right)\right] + &\\
+ \sum_{r\in\mathcal{R}^{\mathsf{Ord}}} \sum_{j=1}^{n_i} I^r_{i,j} \log\left(p_{Y^r_{i,j}-1} - p_{Y^r_{i,j}}\right) + &\\
+ \sum_{r\in\mathcal{R}^{\mathsf{Cat}}} \sum_{j=1}^{n_i} I^r_{i,j} \log\left[\eta^{r,(g)}_{i,j,Y^r_{i,j}} - \log\left(1 + \sum_{k=1}^{K^r-1} \exp\left\{\eta^{r,(g)}_{i,j,k}\right\}\right)\right] &
\end{aligned} \tag{7.8}$$

up to an additive constant common to all clusters $g = 1, \dots, G$. From the similarity to (6.11) we immediately see that h_g is twice differentiable function, gradient and negative Hess matrix of which can be evaluated in the same way as in Section 6.2.3.

The rest of this section is dedicated to numerical approximation of the integral of shape $\int \exp\{h_g(\boldsymbol{b}_i)\}\,\mathrm{d}\boldsymbol{b}_i$, which is achieved by the methodology of *Laplacian approximation* or *adaptive Gaussian quadrature* (AGQ) both summarized, for example, by Pinheiro and Chao (2006). Since the primary goal is to approximate the probabilities $u_{i,g}(\boldsymbol{\theta})$, we often ignore unnecessary multiplicative constants in the process. However, these constants may be important for evaluation of the overall marginal pdf for observed outcomes (3.7), which is needed, for example, to evaluate the contribution to the deviance $D_i^G(\boldsymbol{\theta}; \mathbb{Y}_i^{\mathsf{obs}}, \mathcal{C}_i)$. Keep in mind that the probabilities $(u_{\mathsf{new},g}(\boldsymbol{\theta}))$ could be evaluated even for a newly observed unit as long as the observed outcomes $\mathbb{Y}^{\mathsf{obs}}_{\mathsf{new}}$ and covariates $\mathcal{C}_{\mathsf{new}}$ are at disposal.

Laplacian approximation

Laplacian approximation uses Taylor expansion of function h_g at its maximum $\widehat{\boldsymbol{b}}_i^{(g)}$ that can be found iteratively by Newton-Raphson method, see Algorithm 6. Unlike its use for Metropolis proposal, here we are actually interested not only in the negative Hess matrix $H^{(g)} = -\left[\left.\frac{\partial^2 h_g(\boldsymbol{b}_i)}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^\top}\right|_{\boldsymbol{b}_i = \widehat{\boldsymbol{b}}_i^{(g)}}\right]$ but also in $\widehat{\boldsymbol{b}}_i^{(g)}$ maximizing the function h_g and the value $h_g\left(\widehat{\boldsymbol{b}}_i^{(g)}\right)$.

Since $\widehat{\boldsymbol{b}}_i^{(g)}$ maximizes function h_g, the term corresponding to the first derivative within the Taylor expansion vanishes and we obtain the following approximation

$$h_g(\boldsymbol{b}_i) \approx h_g\left(\widehat{\boldsymbol{b}}_i^{(g)}\right) - \frac{1}{2}\left(\widehat{\boldsymbol{b}}_i^{(g)} - \boldsymbol{b}_i\right)^\top H^{(g)} \left(\widehat{\boldsymbol{b}}_i^{(g)} - \boldsymbol{b}_i\right).$$

Then,

$$\int \exp h_g(\boldsymbol{b}_i)\,\mathrm{d}\boldsymbol{b}_i \approx \exp\left\{h_g\left(\widehat{\boldsymbol{b}}_i^{(g)}\right)\right\} \int \exp\left\{-\frac{1}{2}\left(\widehat{\boldsymbol{b}}_i^{(g)} - \boldsymbol{b}_i\right)^\top H^{(g)} \left(\widehat{\boldsymbol{b}}_i^{(g)} - \boldsymbol{b}_i\right)\right\} \mathrm{d}\boldsymbol{b}_i,$$

where the remaining integral reminds the density of multivariate normal distribution with mean $\widehat{\boldsymbol{b}}_i^{(g)}$ and variance matrix $H^{-(g)} := \left(H^{(g)}\right)^{-1}$. Hence,

$$\int \exp h_g(\boldsymbol{b}_i)\,\mathrm{d}\boldsymbol{b}_i \approx \exp\left\{h_g\left(\widehat{\boldsymbol{b}}_i^{(g)}\right)\right\} \cdot \left|H^{(g)}\right|^{-\frac{1}{2}}$$

up to a multiplicative constant common to all clusters $g = 1, \ldots, G$.

Finally, we can approximate the classification probabilities by

$$u_{i,g}(\boldsymbol{\theta}) \approx \frac{w_g \left|\boldsymbol{\Sigma}^{(g)}\right|^{-\frac{1}{2}} \left|H^{(g)}\right|^{-\frac{1}{2}} \exp\left\{h_g\left(\widehat{\boldsymbol{b}}_i^{(g)}\right)\right\}}{\sum\limits_{g'=1}^{G} w_{g'} \left|\boldsymbol{\Sigma}^{(g')}\right|^{-\frac{1}{2}} \left|H^{(g')}\right|^{-\frac{1}{2}} \exp\left\{h_{g'}\left(\widehat{\boldsymbol{b}}_i^{(g')}\right)\right\}},$$

numerators of which are again calculated on the log-scale first for computational stability.

Adaptive Gaussian quadrature approximation

The Laplacian approach approximates the integral roughly by using the behaviour of function h_g at the single point $\widehat{\boldsymbol{b}}_i^{(g)}$. AGQ also works with approximation of $\exp\{h_g(\boldsymbol{b}_i)\}$ by a density of multivariate normal distribution $\mathsf{N}_{d^{\mathsf{R}}}\left(\widehat{\boldsymbol{b}}_i^{(g)},\, H^{-(g)}\right)$, however, it explores h_g in multiple carefully chosen points and uses them for approximation.

In one-dimensional quadrature based on $\mathsf{N}(0,\, 1)$ distribution, the set of N_{GQ} ideal points $z_j \in \mathbb{R}, j = 1, \ldots, N_{GQ}$ is determined as roots of Hermite polynomial $H_n(x) = (-1)^n e^{x^2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} e^{-x^2}$ with $n = N_{GQ}$. Each root has its own weight $v_j = 2^{n-1} n! \sqrt{\pi} / n^2 \left[H_{n-1}(z_j)\right]^2$. Then, the approximation of the integral of exponentiated univariate function h is determined by

$$\int \exp\{h(x)\}\,\mathrm{d}x = \int \exp\left\{h(x) + x^2\right\} e^{-x^2}\,\mathrm{d}x \approx \sum_{j=1}^{N_{GQ}} w_j \exp\left\{h(z_j) + z_j^2\right\}.$$

In more dimensions (d^{R} in our case), we create vectors $\boldsymbol{z_j} := (z_{j_1}, \ldots, z_{j_{d^{\mathsf{R}}}})$, elements of which are the Hermite polynomial roots indexed by $\boldsymbol{j} = (j_1, \ldots, j_{d^{\mathsf{R}}}) \in \{1, \ldots, N_{GQ}\}^{d^{\mathsf{R}}}$. Vectors $\boldsymbol{z_j}$ of roots are ideal for the standardized case $\mathsf{N}_{d^{\mathsf{R}}}(\boldsymbol{0},\, \mathbb{I})$ and, therefore, are scaled by the inverse of Cholesky triangle $\left(H^{(g)}\right)^{-\frac{1}{2}}$ and shifted by $\widehat{\boldsymbol{b}}_i^{(g)}$ to obtain vectors $\widetilde{\boldsymbol{b}}_{i,\boldsymbol{j}}^{(g)} = \widehat{\boldsymbol{b}}_i^{(g)} + \left(H^{(g)}\right)^{-\frac{1}{2}} \boldsymbol{z_j}$ ideal for $\mathsf{N}_{d^{\mathsf{R}}}\left(\widehat{\boldsymbol{b}}_i^{(g)},\, H^{-(g)}\right)$. Further, we incorporate the norm of $\boldsymbol{z_j}$ into the weights $W_{\boldsymbol{j}} = \exp\left\{\|\boldsymbol{z_j}\|^2\right\} \prod\limits_{l=1}^{d^{\mathsf{R}}} v_{j_l}$. Then, the integral can be approximated by

$$\int \exp\left\{h_g(\boldsymbol{b}_i)\right\}\,\mathrm{d}\boldsymbol{b}_i \approx (2\pi)^{\frac{d^{\mathsf{R}}}{2}} \left|H^{(g)}\right|^{-\frac{1}{2}} \sum_{\boldsymbol{j}} \exp\left\{h_g\left(\widetilde{\boldsymbol{b}}_{i,\boldsymbol{j}}^{(g)}\right)\right\} W_{\boldsymbol{j}}$$

and the classification probabilities by

$$u_{i,g}(\boldsymbol{\theta}) \approx \frac{w_g \left|\boldsymbol{\Sigma}^{(g)}\right|^{-\frac{1}{2}} \left|H^{(g)}\right|^{-\frac{1}{2}} \sum\limits_{\boldsymbol{j}} \exp\left\{h_g\left(\widetilde{\boldsymbol{b}}_{i,\boldsymbol{j}}^{(g)}\right)\right\} W_{\boldsymbol{j}}}{\sum\limits_{g'=1}^{G} w_{g'} \left|\boldsymbol{\Sigma}^{(g')}\right|^{-\frac{1}{2}} \left|H^{(g')}\right|^{-\frac{1}{2}} \sum\limits_{\boldsymbol{j}} \exp\left\{h_{g'}\left(\widetilde{\boldsymbol{b}}_{i,\boldsymbol{j}}^{(g')}\right)\right\} W_{\boldsymbol{j}}}.$$

Note that case $N_{GQ} = 1$ reduces to Laplacian approximation as $z_1 = 0$ and $w_j = 1$ in such case. In applications, we recommend to use rather low value of N_{GQ} since there are $N_{GQ}^{d^{\mathsf{R}}}$ summands to be evaluated, which can be very expensive to compute.

7.4 Basic numerical algorithms

We dedicate the following section to a few basic numerical algorithms referenced during the thesis. These algorithms are broadly known, yet, we present them in the form used for implementation of our methodology.

Cholesky decomposition

When the full-conditional distribution is evaluated or a new value is proposed via Metropolis step, we often work with real symmetric positive-definite matrices. André-Luis Cholesky proposed a decomposition into the product of a lower triangular matrix and its transpose, which is roughly twice as efficient as the LU decomposition for solving a system of linear equations. We provide a pseudocode for the procedure in Algorithm 5. Beside that we use it for inverting variance matrices and sampling from multivariate normal distribution.

If we wish to sample $\boldsymbol{X} \sim \mathsf{N}_d\left(\mathbb{V}^{-1}\boldsymbol{\mu},\, \mathbb{V}^{-1}\right)$ (Sections 5.1.6, 5.1.9) and have the Cholesky decomposition $\mathbb{V} = \mathbb{C}\mathbb{C}^\top$, where $\mathbb{C}$ is lower triangular matrix, then it is sufficient to sample $\boldsymbol{Y} \sim \mathsf{N}_d\left(\mathbb{C}^{-1}\boldsymbol{\mu},\, \mathbb{I}\right)$ and rescale it by $\left(\mathbb{C}^\top\right)^{-1}$ since

$$\left(\mathbb{C}^\top\right)^{-1}\boldsymbol{Y} \sim \mathsf{N}_d\left(\left(\mathbb{C}^\top\right)^{-1}\mathbb{C}^{-1}\boldsymbol{\mu},\, \left(\mathbb{C}^\top\right)^{-1}\mathbb{C}^{-1}\right) = \mathsf{N}_d\left(\mathbb{V}^{-1}\boldsymbol{\mu},\, \mathbb{V}^{-1}\right).$$

The initial mean value $\mathbb{C}^{-1}\boldsymbol{\mu}$ can be found simultaneously with the Cholesky decomposition $\mathbb{V} = \mathbb{C}\mathbb{C}^\top$ as a solution to $\mathbb{C}\boldsymbol{x} \overset{!}{=} \boldsymbol{\mu}$, which is done by Algorithm 5. Then, once the Gaussian noise is added to create $\boldsymbol{Y}$, the later scaling by $\left(\mathbb{C}^\top\right)^{-1}$ is a simple back-solving procedure for the system $\mathbb{C}^\top\boldsymbol{X} \overset{!}{=} \boldsymbol{Y}$. Had the noise not been added, we would obtain the mean value of the target distribution $\mathbb{V}^{-1}\boldsymbol{\mu}$ (a solution to $\mathbb{V}\boldsymbol{x} \overset{!}{=} \boldsymbol{\mu}$ in general).

Algorithm 5 Cholesky decomposition $A = \mathbb{C}\mathbb{C}^\top$ and solution to $\mathbb{C}\boldsymbol{x} \overset{!}{=} \boldsymbol{b}$

```
procedure CholeskySolve($A, \boldsymbol{b}$)
                              ▷ symmetric positive-definite matrix $A$ of dimension $d$
                                           ▷ right-hand side $\boldsymbol{b}$ of dimension $d$
    for $i$ in $1 : d$ do
                                                              ▷ diagonal element first
        $\mathbb{C}[i,i] := A[i,i]$;
        for $k$ in $1 : i$ do
            $\mathbb{C}[i,i] \mathrel{-}= A[k,i] \cdot A[k,i]$;
        end for

        if $\mathbb{C}[i,i] \leq 0$ then
            Error: $A$ is not positive-definite matrix!
        end if
        $\mathbb{C}[i,i] := \sqrt{\mathbb{C}[i,i]}$;
                                                 ▷ forward-solving $\mathbb{C}\boldsymbol{x} \overset{!}{=} \boldsymbol{b}$
        $\boldsymbol{x}[i] = \boldsymbol{b}[i]$;
        for $l$ in $1 : i$ do
            $\boldsymbol{x}[i] \mathrel{-}= \boldsymbol{x}[l] \cdot \mathbb{C}[i,l]$;
        end for
        $\boldsymbol{x}[i] \mathrel{/}= \mathbb{C}[i,i]$;
                                              ▷ remaining elements of column $i$
        for $j$ in $(i+1) : d$ do
            $\mathbb{C}[j,i] := A[j,i]$;
            for $k$ in $1 : i$ do
                $\mathbb{C}[j,i] \mathrel{-}= \mathbb{C}[i,k] \cdot \mathbb{C}[j,k]$;
            end for
            $\mathbb{C}[j,i] \mathrel{/}= \mathbb{C}[i,i]$;
        end for
    end for
    return $\mathbb{C}$ and $\boldsymbol{x}$.
end procedure
```

Newton–Raphson algorithm

Newton–Raphson method, named after Isaac Newton and Joseph Raphson, was originally designed to find the roots of a real valued function. Alternatively, we can view its multivariate version as a tool for finding $\widehat{\boldsymbol{\omega}} \in \mathbb{R}^{\kappa}$ maximising twice differentiable function $\ell(\boldsymbol{\omega})$ since the maximum solves $\frac{\partial \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \overset{!}{=} \mathbf{0}$ and negative Hess matrix is positive-definite matrix. We aim to use it primarily for ℓ representing the log-pdfs of the full-conditional distributions in order to find reasonable proposal distribution, see Section 6.2.

Starting from some initial value $\boldsymbol{\omega}_0$, e.g. the maximum from the previous step, we iteratively solve

$$\left[-\frac{\partial^2 \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top} \bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_k} \right] \boldsymbol{s} \overset{!}{=} \frac{\partial \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_k}.$$

to find the direction in which to move from current position $\boldsymbol{\omega}_k$, see Algorithm 6 for details. The procedure is ended when the norm of the step $\boldsymbol{s}$ is below the tolerance level ϵ. This procedure yields $\widehat{\boldsymbol{\omega}}$ as well as the basis for the precision matrix $\boldsymbol{\Omega}^{-1}$ of the incremental distribution.

The notation of the Algorithm 6 is deceptively simple. We use it for $\ell(\boldsymbol{\psi}|\cdots)$ symbolizing the log-pdf of the target (full-conditional) distribution of $\boldsymbol{\psi} \in \boldsymbol{\Psi}$. Hence, the functions Ell, GradEll and HessEll need to adaptively change with the other parameters for every iteration of the MCMC sampler.

Moreover, there have to be added some checking rules to prevent divergence, overshooting or other failures[*]. For example, limit the number of iterations by a maximal count. If it fails to complete in time or diverges then try different starting point. If the problem is encountered too many times for different starting points then keep the current proposal distribution and update it next time.

[*] Although all the negative Hessian matrices are positive-definite (usually due to the contribution of prior distribution in the form of diagonal positive-definite matrix), the task of finding a solution to the system of equations may be ill-conditioned (high condition number). The contribution of the prior distribution may also have a wild impact on the function behaviour near the root, hence, the overshooting and divergence.

Algorithm 6 Newton–Raphson method to maximise $\ell(\boldsymbol{\omega})$

```
Used functions:
Ell($\boldsymbol{\omega}$)                                    ▷ returning value $\ell(\boldsymbol{\omega})$
GradEll($\boldsymbol{\omega}$)                                ▷ returning gradient $\frac{\partial \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}$
HessEll($\boldsymbol{\omega}$)                   ▷ returning negative Hess matrix $-\frac{\partial^2 \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top}$
CholeskySolve($A, \boldsymbol{b}$) ▷ returning $\mathbb{C}$ from $A = \mathbb{C}\mathbb{C}^\top$ and solution to $\mathbb{C}\boldsymbol{x} \overset{!}{=} \boldsymbol{b}$
BackSolve($\mathbb{C}$, $\boldsymbol{b}$)  ▷ returning solution to $\mathbb{C}^\top \boldsymbol{x} \overset{!}{=} \boldsymbol{b}$ with $\mathbb{C}$ upper triangular

procedure NewtonRaphson($\boldsymbol{\omega}_0, \epsilon$)          ▷ starting point and tolerance
    $\|s\| \leftarrow \infty$;
    $k \leftarrow 0$;
    while $\|s\| > \epsilon$ do
        $\boldsymbol{g} \leftarrow$ GradEll($\boldsymbol{\omega}_k$);
        $H \leftarrow$ HessEll($\boldsymbol{\omega}_k$);
        $\mathbb{C}, \boldsymbol{b} \leftarrow$ CholeskySolve($H, \boldsymbol{g}$);
        $\boldsymbol{s} \leftarrow$ BackSolve($\mathbb{C}, \boldsymbol{b}$);                ▷ solution to $H\boldsymbol{x} \overset{!}{=} \boldsymbol{g}$
        $\boldsymbol{\omega}_{k+1} := \boldsymbol{\omega}_k + \boldsymbol{s}$;
        compute $\|\boldsymbol{s}\|$;
        $k \leftarrow k + 1$;
    end while
    $\widehat{\boldsymbol{\omega}} := \boldsymbol{\omega}_k$;
    return $\widehat{\boldsymbol{\omega}}$ or even Ell($\widehat{\boldsymbol{\omega}}$), GradEll($\widehat{\boldsymbol{\omega}}$), HessEll($\widehat{\boldsymbol{\omega}}$).
end procedure
```

k-means clustering

As discussed in Chapter 3, multivariate data often tend to form into clusters without the prior knowledge of the true allocation of each data point. The goal of unsupervised clustering is to learn the ideal partition from the data. Favourite distance-based solution is the k-means algorithm dividing the data into k clusters concentrated around k centroids. The first fundamental ideas for the k-means clustering appeared in 1956 and one year later Stuart Lloyd proposed the standard naive algorithm to obtain a locally (not globally) optimal partition.

Let us assume that the rows of matrix $\mathbb{X} \in \mathbb{R}^{n\times d}$ are the independent d-variate data points $\boldsymbol{x}_i, i = 1, \ldots, n$. The naive algorithm starts with k initial centroids $\boldsymbol{\mu}_1^0, \ldots, \boldsymbol{\mu}_k^0 \in \mathbb{R}^d$, perhaps, cluster means of a random partition. And then iterate the following two steps (and increase the counter j) until the convergence is met (no change in the partition):

1) Assign each data point i to the group with the closest centroid (in terms of Euclidean distance $\| \bullet - \star \|$):

$$U_i^j = \operatorname*{arg\,min}_{g\in\{1,\ldots,k\}} \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_g^{j-1} \right\|.$$

2) Compute the means in each cluster and use them as the next centroids:

$$\mathcal{U}_g^j := \{i : U_i^j = g, i = 1, \ldots, n\} \quad \text{and} \quad \boldsymbol{\mu}_g^j := \frac{1}{\left|\mathcal{U}_g^j\right|} \sum_{i\in\mathcal{U}_g^j} \boldsymbol{x}_i.$$

The naive algorithm suffers from slow convergence since it spends a lot of time computing the distances of points from the centroids, which in most cases is unnecessary since the majority of data points after few initial iterations stay within the same cluster. Hence, Hartigan and Wong (1979) proposed a completely different method for updating the partition. They base their update step on the individual cost $\varphi(\mathcal{U}_g)$ of a cluster $\mathcal{U}_g \subset \{1, \ldots, n\}$ defined by

$$\varphi(\mathcal{U}_g) := \sum_{i\in\mathcal{U}_g} (\boldsymbol{x}_i - \boldsymbol{\mu}_g)^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_g)$$

where $\boldsymbol{\mu}_g$ is the centre of the cluster g. Starting from an initial partition, a point maximizing the given criterium Δ switches to a different cluster at each iteration until no optimal change is available, see Algorithm 7 for details.

For efficient computation of $\Delta(i, g)$ it is crucial to effectively evaluate φ of a cluster after addition or deletion of a single data point:

$$\begin{aligned}
\overline{\boldsymbol{x}}_n &:= \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i, & \varphi_n &:= \sum_{i=1}^n \left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n\right)^\top \left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n\right), \\
\overline{\boldsymbol{x}}_{n-1} &= \frac{n}{n-1}\overline{\boldsymbol{x}}_n - \frac{\boldsymbol{x}_n}{n-1}, & \varphi_{n-1} &= \varphi_n - \frac{n}{n-1}\left(\boldsymbol{x}_n - \overline{\boldsymbol{x}}_n\right)^\top \left(\boldsymbol{x}_n - \overline{\boldsymbol{x}}_n\right), \\
\overline{\boldsymbol{x}}_{n+1} &= \frac{n}{n+1}\overline{\boldsymbol{x}}_n + \frac{\boldsymbol{x}_{n+1}}{n+1}, & \varphi_{n+1} &= \varphi_n + \frac{n}{n+1}\left(\boldsymbol{x}_{n+1} - \overline{\boldsymbol{x}}_n\right)^\top \left(\boldsymbol{x}_{n+1} - \overline{\boldsymbol{x}}_n\right).
\end{aligned}$$

The values $\varphi\left(\mathcal{U}_g^j\right)$ are computed for the initial partition and then updated only for the changed clusters, no need to compute them repeatedly. The same holds even for $\Delta(i,g)$ where $i \neq \widehat{i}$ and $g \neq \widehat{g}$ (optimal pair from the previous step), which is where the computational time is saved compared to the naive algorithm.

Nevertheless, these changes only speed up the calculation. The algorithm still ends up with locally optimal solution, the globally optimal solution is not guaranteed. Initialization by different partitions may help to improve the final solution, however, the global optimality cannot be assured.

We utilize the implemented function `kmeans` from base R where the version by Hartigan and Wong (1979) is the default choice.

Algorithm 7 k-means clustering by Hartigan and Wong (1979)

```
procedure kMeansHW($\mathbb{X}, k$)    ▷ $n \times d$ data matrix and number of clusters
    Initialize the partition of the data into $k$ clusters $\mathcal{U}_1^0, \ldots, \mathcal{U}_k^0$;
                ▷ e.g. $U_i^0 \overset{\text{iid}}{\sim} \mathsf{Unif}\,\{1,\ldots,k\}$ and $\mathcal{U}_g^0 := \{i : U_i^0 = g, i = 1,\ldots,n\}$
    $j = 0$;
    repeat
        for $i$ in $1:n$ and $g$ in $1:k$ do    ▷ only some pairs require the update
            $\Delta(i,g) := \varphi\left(\mathcal{U}^j_{U_i^j}\right) + \varphi\left(\mathcal{U}_g^j\right) - \varphi\left(\mathcal{U}^j_{U_i^j} \setminus \{i\}\right) - \varphi\left(\mathcal{U}_g^j \cup \{i\}\right)$;
                    ▷ the change in the cost when $i$ switches to cluster $g$
        end for
        $\left(\widehat{i}, \widehat{g}\right) := \operatorname*{arg\,max}_{i \in \{1,\ldots,n\},\, g \in \{1,\ldots,k\}} \Delta(i,g)$;    ▷ pair maximizing the change
        $\widetilde{g} := U_{\widehat{i}}^j$;    ▷ cluster where $\widehat{i}$ currently belongs to

        if $\Delta\left(\widehat{i}, \widehat{g}\right) < 0$ then
            leave the repeat cycle;
                ▷ no better partition achieved by switching a single data point
        else
            $U_i^{j+1} := U_i^j$ and $\mathcal{U}_g^{j+1} := \mathcal{U}_g^j$;    ▷ other clusters remain the same
            $U_{\widehat{i}}^{j+1} := \widehat{g}$, $\mathcal{U}_{\widetilde{g}}^{j+1} := \mathcal{U}_{\widetilde{g}}^j \setminus \left\{\widehat{i}\right\}$ and $\mathcal{U}_{\widehat{g}}^{j+1} := \mathcal{U}_{\widehat{g}}^j \cup \left\{\widehat{i}\right\}$;
                    ▷ switch the data point $\widehat{i}$ to cluster $\widehat{g}$
        end if
        $j \leftarrow j + 1$;
    until $\Delta\left(\widehat{i}, \widehat{g}\right) < 0$    ▷ no change in the partition
return partition $\{\mathcal{U}_1^j, \ldots, \mathcal{U}_k^j\}$ and means $\boldsymbol{\mu}_1^j, \ldots, \boldsymbol{\mu}_k^j$.
end procedure
```

8. Analysis of the EU-SILC

In Chapter 1 we introduced the EU-SILC dataset that gathers data from European households regarding their financial situation and quality of life. We have presented the highly correlated outcomes of interest as well as covariates potentially effecting these outcomes.

Within this chapter we apply only the most recent *GLMM-based* model to this dataset since it provides several advantages. First of all, it is able to work with *Yes / No / No cannot* outcomes as general categorical and not necessarily ordinal. Then, we do not have to exclude several households for not delivering some of the outcomes. Moreover, our specification of the model predictor can be tailored specifically to fit our goals.

And that is to model the chosen outcomes of different nature jointly by one statistical model, which can evaluate the effects of covariates. Since we hypothesise that the economical crisis had serious impact on the financial situation of the households (either positive, negative or none), we are particularly interested in the exploration of evolution in time. We expect to find clusters of different trends depicting the capabilities to cope with such a crisis.

8.1 The model setting

First, we had to adequately pair households from the longitudinal dataset to the ones in the cross-sectional dataset to obtain outcomes and covariates not included within the longitudinal dataset, such as the household type. The data from each year are stored in separate files, hence, there would be multiple repeated data rows, had only the records from the year the household is observed for the last time not been used. The household module (H-file) is the source of the outcomes measured at household-level, however, some of the covariates (presence of a student, a baby, highest ISCED achieved) had to be aggregated from the data measured at person-level (P and R-file) within the household. After merge and aggregation the households not observed for the whole period of four consecutive years were disregarded. The same fate befell households with missing covariate values (missing values of the outcomes were allowed since the implementation of the *GLMM-based* model overcomes this issue). After this data pre-processing we were left with $n = 27\,386$ Czech households observed for exactly $n_i = 4$ consecutive years between the years 2005 and 2020.

All 8 outcomes introduced in Section 1.3.2 (two for each type excluding the count type) were modelled jointly through our *GLMM-based* model. The numeric outcomes (*Equivalised total disposable income* and *Lowest income to make ends meet*) were log-transformed to better fit the assumption of normal distribution behind. In some instances of negative disposable income, we had to replace the log-value by zero. Other categorical outcomes were transformed to values $0, 1, \ldots, K^r - 1$ to fit our model assumptions, where the zero level corresponds to the baseline categories. For the binary outcomes $0 =$ *No*, for the general categorical $0 =$ *Cannot* and for the ordinal outcomes zero corresponds to the poorest category (*with great difficulty* and *a heavy financial burden*). Hence, all the outcome values are aligned to have the same interpretation; the higher value the more rich the household is considered.

To join these 8 outcomes we suppose single random intercept term for each of them resulting in the random effects vector $\boldsymbol{b}_i \in \mathbb{R}^8$. The random intercepts for general categorical outcomes are not specific to each category level and compare only the baseline category *Cannot* with the rest. Hence, the relationships among outcomes are described by the correlations hidden in $0 < \boldsymbol{\Sigma} \in \mathbb{R}^{8\times 8}$ among the random intercepts. Since the categories are ordered and aligned with the income variables, only positive correlations are expected to be found.

The fixed effects structure is far richer. For all outcomes of interest it consists of all the covariates listed in Section 1.3.3 in a proper parametrization. Since the economical crisis can have both immediate and long-lasting effects, the parametrization of time ($T_{i,j}$) needs to be very flexible. Hence, we use quadratic B-spline (S_k) parametrization with 5 equidistant knots in the interval $[0, 16)$ (including the boundary ones), which takes five β parameters denoted by $\beta_1, \dots, \beta_5$ when excluding the random intercept. The *Equivalised household size* (W as weight) enters the predictor in a linear form (with β_W) shifted by 1 corresponding to a single-adult household. Other covariates are of categorical nature and each non-baseline category has its own effect. For example, the *Level of urbanisation* has effects $\beta_{U2}, \beta_{U3}, \beta_{U4}$ comparing other areas with the baseline rural area. Similarly, we denote the effects for *Presence of a baby, a student* by β_B, β_S; for education by β_{E2}, β_{E3}; for the *Dwelling type* by $\beta_{D2}, \dots, \beta_{D5}$ and for the *Household type* by $\beta_{H6}, \dots, \beta_{H13}$ with type 5 (one person household) as baseline. Denoting the covariates analogously the predictor becomes:

$$\begin{aligned}\eta_{i,j}^{r} = b_i^r + \beta_{0,r} + \sum_{k=1}^{5} \beta_{k,r} S_k(t_{i,j}) + \beta_{W,r}(W_{i,j} - 1) + \sum_{k=2}^{4} \beta_{Uk,r} \mathbb{1}_{(U_{i,j}=k)} + \beta_{B,r} B_{i,j} + \\ + \beta_{S,r} S_{i,j} + \sum_{k=2}^{3} \beta_{Ek,r} \mathbb{1}_{(E_{i,j}=k)} + \sum_{k=2}^{5} \beta_{Dk,r} \mathbb{1}_{(D_{i,j}=k)} + \sum_{k=6}^{13} \beta_{Hk,r} \mathbb{1}_{(H_{i,j}=k)} \quad (8.1)\end{aligned}$$

for any outcome $r \in \mathcal{R}$, household $i = 1, \dots, n$ and observation $j = 1, 2, 3, 4$.

However, the primary goal is to divide the households into groups differing in the evolution in time (different types of impact of the economical crisis). We could simply assume all the unknown parameters to be cluster-specific, however, that would result in enormous amount of parameters and describing them all would be exhausting. Hence, we make cluster-specific only those parameters which parametrize the effect of time: the intercept term β_0, ordered intercepts for ordinal outcomes $\boldsymbol{c}$ and primarily the effects of spline bases $\beta_1, \dots, \beta_5$. Other model parameters (the rest of fixed effects, variance matrix $\boldsymbol{\Sigma}$) are considered to be common to all groups with the exception of parameter $\boldsymbol{\tau}$ describing the variability of error terms of numeric outcomes for which we try both settings. The predictor (8.1) for group g becomes

$$\begin{aligned}\eta_{i,j}^{r,(g)} = b_i^r + \beta_{0,r}^{(g)} + \sum_{k=1}^{5} \beta_{k,r}^{(g)} S_k(t_{i,j}) + \beta_{W,r}(W_{i,j} - 1) + \sum_{k=2}^{4} \beta_{Uk,r} \mathbb{1}_{(U_{i,j}=k)} + \beta_{B,r} B_{i,j} + \\ + \beta_{S,r} S_{i,j} + \sum_{k=2}^{3} \beta_{Ek,r} \mathbb{1}_{(E_{i,j}=k)} + \sum_{k=2}^{5} \beta_{Dk,r} \mathbb{1}_{(D_{i,j}=k)} + \sum_{k=6}^{13} \beta_{Hk,r} \mathbb{1}_{(H_{i,j}=k)}. \quad (8.2)\end{aligned}$$

Since the number of groups is not known in advance, we employ the sparse finite mixtures by setting the hyperparameters of the prior (4.12) to $a_e = 1$ and

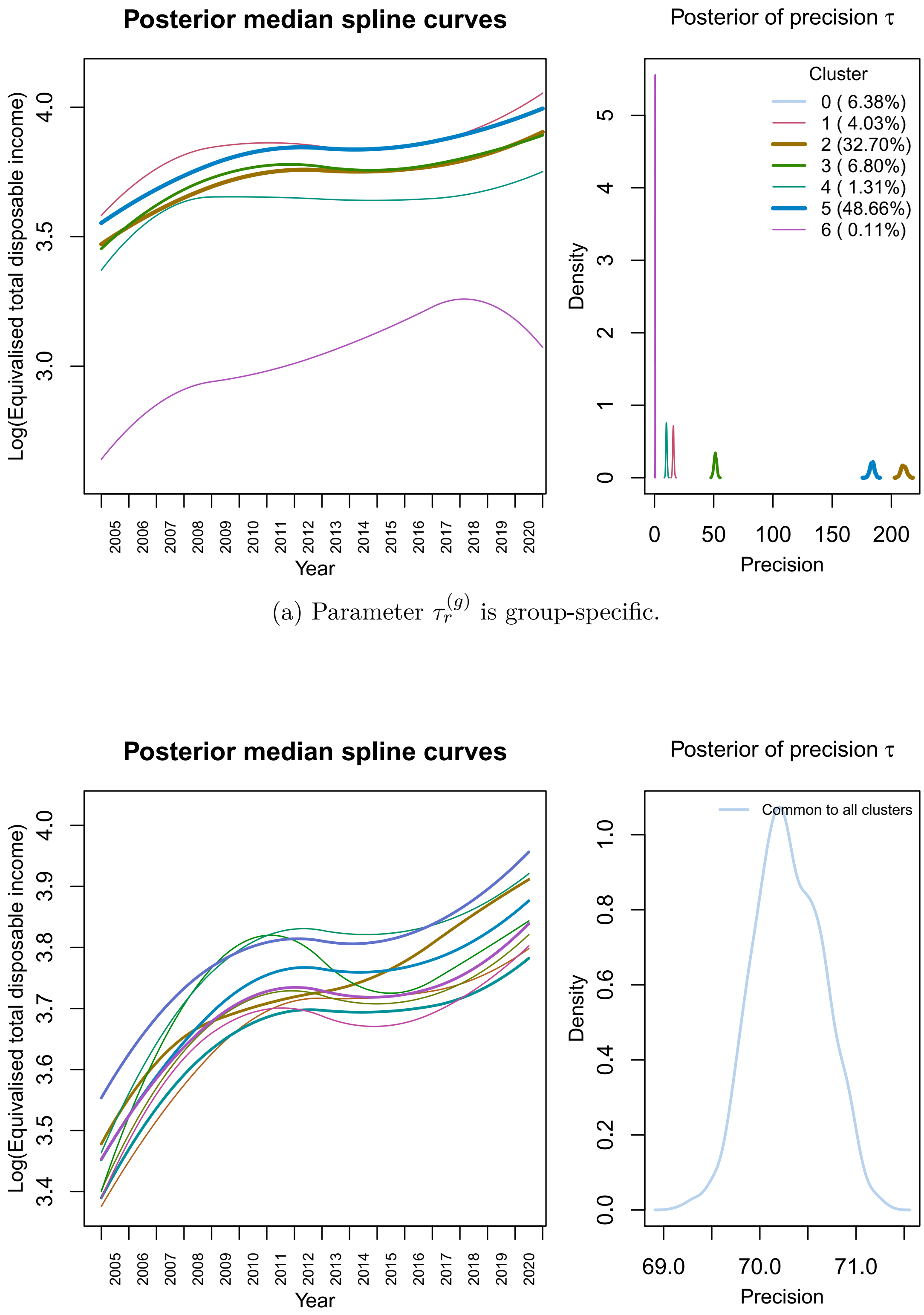
$b_e = 100$ and supposing $G_{\mathsf{max}} = 20$ underlying clusters, some of which may end up empty. Only the remaining $\widehat{G}_+$ non-empty clusters will be considered and interpreted. The rest of hyperparameters are set up to induce prior distribution of unit variance rather then fully uninformative priors (of high variance) since it encourages the shrinkage and by doing so cuts off the redundant mixture components, see the discussion in Section 6.3. Several thousands of burn-in draws were sampled prior the final inference based on $M = 10\,000$ posterior draws. All the states of the final chains used for inference suggested the same number of non-empty clusters $\widehat{G}_+$ and quick overview of the traceplots did not indicate any *label-switching* problems, hence, post-processing procedure (Algorithm 3) was unnecessary.

8.2 Results

First, we address the problem of unknown number of underlying clusters. Surprisingly, our method provides two contradictory answers depending on the cluster-specificity of the parameter $\boldsymbol{\tau}$. On one hand, we have discovered $\widehat{G}_+ = 6$ groups when each numeric outcome $r \in \mathcal{R}^{\mathsf{Num}}$ has its own precision parameter $\tau_r^{(g)}$. On the other hand, there are $\widehat{G}_+ = 11$ non-empty clusters when there is only one τ_r common to all clusters.

Figure 8.1 compares these two options in terms of the most important outcome - *Equivalised total disposable income*. The plots on the left show posterior median curves for the baseline households in each cluster. We immediately see that the obtained results are far from the expected. The curves do not differ that much in the actual shape but more in the intercept that shifts them. It seems that the within variability of the households is much more decisive for the clustering than the trend itself. When all the clusters have to keep the same variability around the curves there is much more room to focus on the trend itself. However, since we observe one household for a limited time window while the overall time span is much larger, it is forced to stay with a general trend. Moreover, the *Equivalised total disposable income* is not the only modelled outcome, the other outcomes surely contribute to the partition in some way. If some splines in Figure 8.1b coincide, there exists some outcome for which the two spline curves of evolution differ more. And that is, perhaps, the reason behind the large number of clusters since many combinations are viable. Once the precisions gain their freedom, the model notices that it is much more suitable to divide the households based on the income changes from year to year. Hence, majority of households evolve steadily (high precision $\tau_r^{(g)}$), while the remaining ones encounter sudden up-and-down shocks (as seen later in Figure 8.2). The last violet cluster is even more special since it covers around 30 households of temporary negative total disposable income which were imputed by 0 instead of undefined log-value. This jump to zero is accompanied by remarkably low precision $\tau_r^{(6)}$. Hence, this cluster should be considered as a cluster of outliers. For the simplicity of the six-cluster solution under group-specific $\tau_r^{(g)}$, we chose this setting for the further analysis.

Next, we explore the relationships among the outcomes. To be precise, covariance matrix $\boldsymbol{\Sigma}$ describes directly only the relationships among the random intercepts, which manifest into the marginal associations between the outcomes once the random effects are integrated out (see Section 2.4). The matrix $\boldsymbol{\Sigma}$ can be



(a) Parameter $\tau_r^{(g)}$ is group-specific.

(b) Parameter τ_r is common to all groups.

Figure 8.1: EU-SILC dataset. Estimated spline curves of *Equivalised total disposable income* based on posterior medians of the $\boldsymbol{\beta}^{(g)}$ coefficients in non-empty clusters for a baseline household (left) and the posterior distribution of the precision parameter τ_r (right).

decomposed into the correlation matrix squeezed between the diagonal matrices with standard deviations, posterior medians of which are

$$(0.11, \quad 0.12, \quad 3.36, \quad 3.46, \quad 2.88, \quad 2.94, \quad 3.05, \quad 3.35)^\top .$$

The posterior medians of the correlations take the following form:

$$\begin{pmatrix}
1.00 & 0.37 & 0.48 & 0.45 & 0.45 & 0.33 & 0.37 & 0.34 \\
0.37 & 1.00 & 0.03 & -0.03 & -0.06 & -0.04 & 0.05 & 0.04 \\
0.48 & 0.03 & 1.00 & 0.80 & 0.74 & 0.58 & 0.54 & 0.48 \\
0.45 & -0.03 & 0.80 & 1.00 & 0.74 & 0.56 & 0.53 & 0.47 \\
0.45 & -0.06 & 0.74 & 0.74 & 1.00 & 0.80 & 0.60 & 0.52 \\
0.33 & -0.04 & 0.58 & 0.56 & 0.80 & 1.00 & 0.50 & 0.45 \\
0.37 & 0.05 & 0.54 & 0.53 & 0.60 & 0.50 & 1.00 & 0.74 \\
0.34 & 0.04 & 0.48 & 0.47 & 0.52 & 0.45 & 0.74 & 1.00
\end{pmatrix}$$

where each row corresponds to the random intercept of the corresponding outcome (in order in which the outcomes are introduced in Section 1.3.2, from numeric, binary, ordinal to general categorical outcomes). As expected, the correlations are mostly positive and definitely not negligible, which confirms the positive association among the considered outcomes. The highest observed correlations are understandably between the binary outcomes (*Affordability of a one week holiday* and *Afford to pay for unexpected expenses*) and the ordinal outcomes (*Ability to make ends meet* and *Financial burden of the total housing cost*). On the other hand, the *Lowest income to make ends meet* (second row) is related only to the *Equivalised total disposable income* (higher income promotes more luxurious, thus, more expensive life style) but barely relates to any other outcome.

The households were classified based on the sampled allocations U_i^m according to the rule (P1); household remained unclassified when the most frequent cluster label did not overcome the 60% threshold. There were 6.38% of unclassified households in total. The purpose of Figure 8.2 is to show the differences among the discovered clusters under group-specific precisions $\tau_r^{(g)}$ across all modelled outcomes. For the numeric outcomes we displayed longitudinal profiles of representatives of each of the clusters (thin lines) and corresponding splines curves (bold lines) that belong to a household of two adults of secondary education with one child (student) living in a detached house in a town (posterior median of the predictor treated as a parametric function). The evolution of categorical outcomes is depicted by the proportions in each cluster and year separately, where the difference between the clusters lies primarily in the magnitude of the proportions and less notably in the evolution over time.

The last (purple) cluster (0.11%) are the outliers in the *Equivalised total disposable income*. The turquoise cluster (1.31%) is also very thin and appears to be the poorest due to the long low-placed slightly decreasing curve in the income, the proportions of the negative categories appear to be the highest among all clusters. The much larger orange cluster (32.70%) resembles the turquoise in the proportions of categorical outcomes but enjoys much higher non-decreasing income. More importantly, the precision is the highest among all clusters (Figure 8.1), which is in accordance with barely changing individual curves. The green cluster (6.80%) follows the same curves as the orange cluster, however, the precision is reduced and the proportions in binary outcomes are more favourable. Moreover, the vast majority of green households possess a computer unlike the orange ones which do not despite being able to afford it. The remaining two clusters both have the highest *Equivalised total disposable income* among them all. The households within the red cluster (4.03%) are, however, more inclined towards the shock jumps between the years due to the precision as low as in the turquoise cluster. On the other hand, the households within the largest (blue) cluster (48.66%) are far more stable (almost as in the orange cluster). The red and the blue cluster also achieve the best ratios of the positive categories.

The effects of time are by (8.2) the only source of the difference among the groups. The remaining covariates were assumed to have the same effect in each cluster. There has been $d_r^{\mathsf{F}} = 20$ fixed effects coefficients per one outcome, which is difficult to interpret in full completeness. Hence, we present only the effects (multiplied by 100 to improve readability) for the log-transformed numeric outcomes *Equivalised total disposable income* and *Lowest income to make ends meet* in Table 8.1. It presents the posterior medians together with 95% ET credible intervals, many of which do not cover zero. This suggests significant effects of the chosen covariates. Naturally, the more urbanized the area where the household is situated the higher income (and the lowest income required) is expected. The size of the household decreases the *Equivalised total disposable income*, while naturally *Lowest income to make ends meet* grows with the size. Households with babies or students have lower disposable income, but presence of a student increases the necessary monthly income. Without any doubt, the higher education level achieved results in much higher income. The differences between dwelling types with respect to the disposable income are negligible (maybe with exception of semi-detached house compared to other types). However, the flats are less demanding on the necessary monthly income compared to a detached house. The composition of the household gives similar results to those seen in Figure 1.7.

Table 8.1: EU-SILC dataset. Posterior medians of the fixed effects of numeric outcomes common to all clusters including 95% equal-tailed credible intervals.

Parameter	Numeric outcome			
	Log(Equivalised total disposable income)		Log(Lowest income to make ends meet)	
$100\beta_{U2}$	1.25	(0.81; 1.55)	0.57	(0.20; 0.88)
$100\beta_{U3}$	2.02	(1.46; 2.40)	1.78	(1.30; 2.20)
$100\beta_{U4}$	7.93	(7.33; 8.71)	10.07	(9.56; 10.59)
$100\beta_{W}$	−6.64	(−7.33; −5.93)	7.49	(6.75; 8.50)
$100\beta_{B}$	−3.88	(−4.21; −3.54)	−1.62	(−1.97; −1.22)
$100\beta_{S}$	−0.92	(−1.22; −0.60)	0.49	(0.20; 0.78)
$100\beta_{E2}$	9.12	(8.51; 9.78)	8.29	(6.77; 8.79)
$100\beta_{E3}$	17.38	(16.77; 18.03)	11.15	(10.11; 11.70)
$100\beta_{D2}$	0.65	(0.24; 1.02)	0.15	(−0.28; 0.63)
$100\beta_{D3}$	0.21	(−0.33; 0.61)	−0.34	(−0.86; 0.06)
$100\beta_{D4}$	0.24	(−0.23; 0.51)	−0.30	(−0.73; 0.00)
$100\beta_{D5}$	1.00	(−0.51; 2.28)	1.78	(0.32; 3.30)
$100\beta_{H6}$	14.68	(14.12; 15.62)	13.90	(13.43; 14.55)
$100\beta_{H7}$	9.65	(8.98; 10.05)	11.97	(11.30; 12.56)
$100\beta_{H8}$	21.42	(20.47; 22.74)	18.55	(17.60; 19.38)
$100\beta_{H9}$	3.11	(2.32; 3.91)	12.23	(11.46; 13.02)
$100\beta_{H10}$	15.24	(14.44; 16.31)	18.60	(17.87; 19.44)
$100\beta_{H11}$	14.22	(13.17; 15.33)	20.49	(19.25; 21.39)
$100\beta_{H12}$	14.39	(13.13; 15.61)	21.28	(19.68; 22.63)
$100\beta_{H13}$	19.79	(18.50; 21.17)	20.06	(18.76; 21.19)

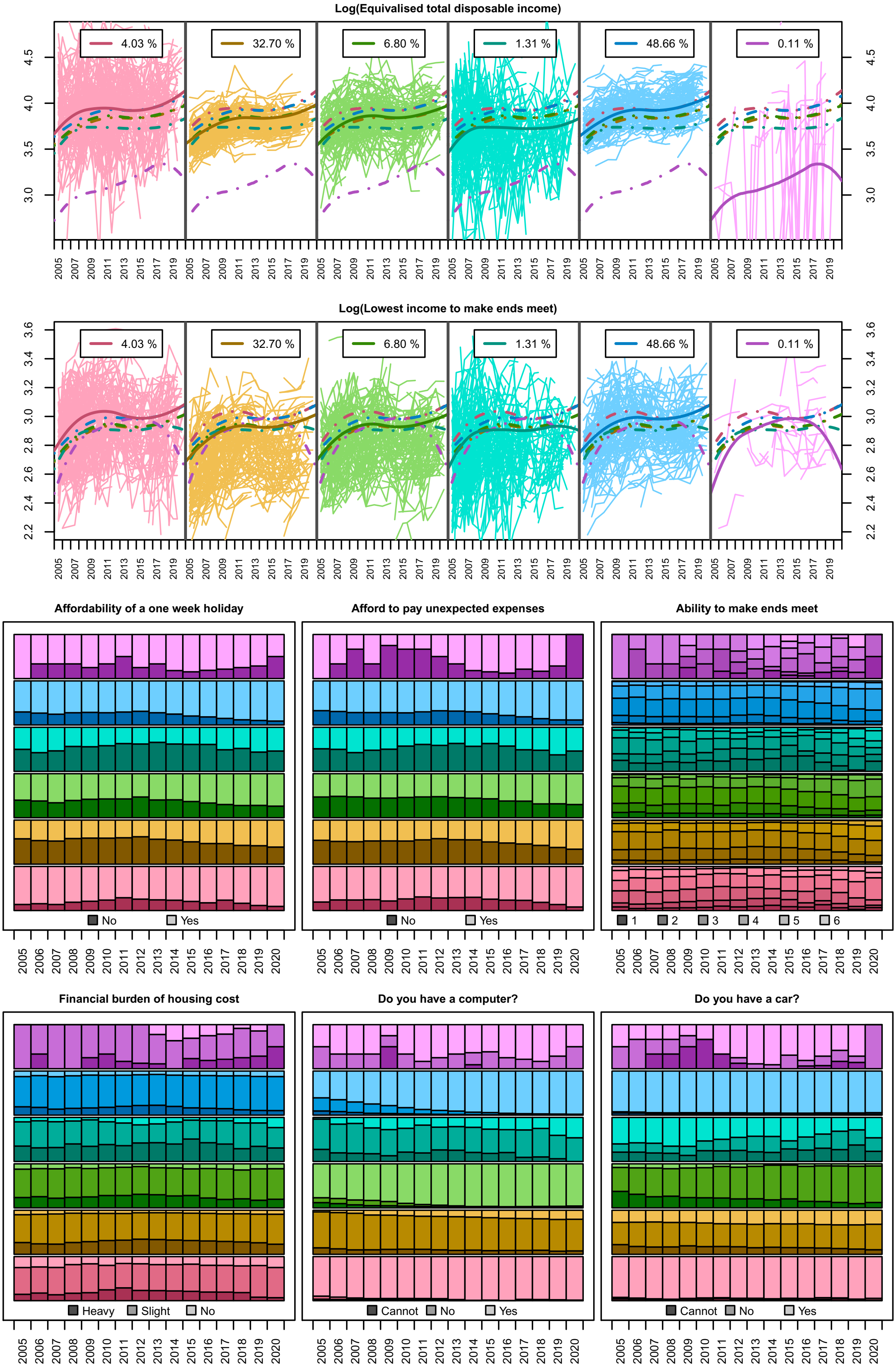


Figure 8.2: EU-SILC dataset, *GLMM-based* model, $G_{\mathsf{max}} = 20$, $\widehat{G}_{+} = 6$. Estimated group-specific spline curves for a typical family by posterior median. Proportions of categorical outcomes wrt time separately in each cluster.

Conclusion

Two classes of statistical models for a set of longitudinal outcomes of diverse nature have been proposed. Combination of potentially highly correlated outcomes was allowed due to joint distribution of random effects of underlying mixed models capturing the main association structure. A model-based clustering framework was adopted to divide the units into groups of different characteristics and, hence, to capture potential heterogeneity within the data. Hierarchical nature of the two models was then exploited within a fully Bayesian approach and MCMC samplers to explore the posterior distribution of the model parameters and characteristics were theoretically justified. MCMC samplers elegantly evade the integration with respect to the latent elements. In case the posterior distribution of classification probabilities or even the model deviance is of interest, we face the unpleasant integrals with a rather expensive numerical approach for evaluation of the likelihood contribution of a single unit. The implementation within the R environment provides the user with full flexibility to specify not only the prior distribution and tuning parameters but most importantly the set of group-specific parameters that distinguish the clusters. The functionality of our methods (ability to estimate the true model parameters and to correctly classify the units) was tested in the simulation studies with relatively positive results. Using only the biomarkers repeatedly observed during the first 910 days we have successfully divided patients from the PBC study into two groups differing in the prognosis for their survival. We have also identified several groups of Czech households differing in the stability of their income, ability to pay for the usual expenses, such as housing costs, and ability to afford luxuries such as a car, a computer or a holiday away from home.

The initial *threshold concept* model summarized in Section 4.4.1 is only able to model numeric, binary and ordinal outcomes. There is a simple reason behind this decision. The threshold concept used here allows us to transfer the binary and ordinal outcomes into the remaining case of a numeric outcome. Hence, the heart of the model is the multivariate normal linear mixed-effects model. Consequently, all the full-conditional distributions fall into the well known distributional families, which allows for direct implementation of the Gibbs sampler. Despite the straightforward implementation, there has occurred a problem of slow convergence of $\boldsymbol{\gamma}$ thresholds to the stationary distribution. The numerical evaluation of the likelihood involved integration of highly-dimensional normal density over a multivariate interval, for which another MCMC-based technique is used. This makes the evaluation of the posterior distribution of the model deviance very expensive. Since the capabilities of the model were still far from our ideal solution, we continued in the research.

We removed the threshold concept and replaced it with the GLMM framework (Section 2.2) which offers plenty of distributional families – logistic regression for binary outcomes, ordinal logit regression for ordinal outcomes and, additionally, multinomial logit regression for general categorical outcomes and log-linear mixed model for count outcomes, which has lead to the so called *GLMM-based* model summarized in Section 4.4.2. This leads no longer to familiar full-conditional distributions of some of the parameters, hence, corresponding steps are replaced with Metropolis proposals. In order to make the proposals more efficient and to spare the user of tuning we have also included a methodology for an automated suggestion of the proposal distribution parameters. Moreover, the implementation accounted for missing outcome values by making them yet another model parameters to be sampled, hence, predictive distributions for the missing values are in the end at disposal. The most important extension lies within the sparse finite mixture which encourages the sampler to abandon the unpromising mixture components only to end up with an optimal number of non-empty components, thus, estimate the number of groups without any prior knowledge. Another benefit is that the fixed part of the predictor has been divided into a group specific part and a part common to all clusters, which provides the user with even higher flexibility in the specification of the differences among clusters. If one settles with a crude Laplace approximation of the likelihood contribution, the exploration of the posterior of classification probabilities and deviance is less time-consuming than for the first model. However, more precise approximation of the multivariate integral with respect to random effects via adaptive Gaussian quadrature could potentially be even more expensive.

There are many possible directions for the future research to further enhance the methodology developed in this thesis. One very straightforward way would be to extend the *GLMM-based* model to any possible combination of distributional family and link functions. Here we assumed only one particular choice for each type of the variable, which is limiting especially for the numeric outcomes. Certainly, there are circumstances where no suitable transformation exists to solve non-normality issues. The separate models would still be possible to join through the joint distribution of random effects given by the covariance matrix $\boldsymbol{\Sigma}$.

The matrix $\boldsymbol{\Sigma}$ also needs a careful attention. One should note that its dimension rises with the number of modelled outcomes and the complexity of the random effects structure. Hence, it may prove useful to abandon the complete generality and replace the completely general design of $\boldsymbol{\Sigma}$ with a commonly used block-structured variance matrix. However, the sampling mechanism for such $\boldsymbol{\Sigma}$ would have to be updated to reflect the implied restriction of the space of all positive-definite matrices.

Moreover, we have examined the properties of our methodology under a rather low number of outcomes so far. Additional work may thus focus on a much higher number of measured outcomes and possibly on an evaluation of their relevance towards clustering using, for example, methods presented by Raftery and Dean (2006). The variable selection process could also be extended to the regression part of the model, which could help the analyst with evaluation of the importance of individual β parameters (or their groups). With some innovative approach we could not only determine the significance of the effect (or groups of effects) but also evaluate the importance with respect to the clustering. However, that would require additional research.

Finally, the whole methodology was implemented as a set of C routines integrated into R (R Core Team, 2022). Researchers interested in the use of our models can access the implementation via Github at `https://github.com/vavrajan/`. Follow the tutorial to learn how to use the functions including the input data preparation and the output data analysis.

Bibliography

EU-SILC Methodological Guidelines. URL https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

M. Aitkin, C. C. Liu, and T. Chadwick. Bayesian model comparison and model averaging for small-area estimation. *The Annals of Applied Statistics*, 3(1): 199–221, 2009. doi: 10.1214/08-aoas205.

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993. doi: 10.2307/2290350.

J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993. doi: 10.2307/2532201.

R. D. Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, 1972. doi: 10.1007/bf02291411.

R. D. Bock and M. Lieberman. Fitting a response model for n dichotomously scored items. *Psychometrika*, 35:179–197, 1970.

C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108644181.

S. Brooks, A. Gelman, G. Jones, and X. Meng. *Handbook for Markov Chain Monte Carlo*. Taylor & Francis, 2nd edition, 2011. ISBN 978-1-4200-7941-8. doi: 10.1201/b10905.

L. Bruckers, G. Molenberghs, P. Drinkenburg, and H. Geys. A clustering algorithm for multivariate longitudinal data. *Journal of Biopharmaceutical Statistics*, 26(4):725–741, 2016.

G. Celeux, O. Martin, and C.Lavergne. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, 5(3):243–267, 2005. doi: 10.1191/1471082X05st096oa.

R. De la Cruz-Mesía, F. A. Quintana, and G. Marshall. Model-based clustering for longitudinal data. *Computational Statistics & Data Analysis*, 52(3):1441–1457, 2008. doi: 10.1016/j.csda.2007.04.005.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

M. J. Denwood. runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, 71(9):1–25, 2016. doi: 10.18637/jss.v071.i09.

E. R. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy. Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 10(1):1–7, 1989. doi: 10.1002/hep.1840100102.

S. Fieuws and G. Verbeke. Joint modelling of multivariate longitudinal profiles: Pitfalls of the random-effects approach. *Statistics in Medicine*, 23:3093–3104, 2004. doi: 10.1002/sim.1885.

S. Fieuws and G. Verbeke. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431, 2006. doi: 10.1111/j.1541-0420.2006.00507.x.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458): 611–631, 2002. doi: 10.1198/016214502760047131.

S. Frühwirth-Schnatter. Panel data analysis: a survey on model-based clustering of time series. *Advances in Data Analysis and Classification*, 5(4):251–280, 2011. doi: 10.1007/s11634-011-0100-0.

S. Frühwirth-Schnatter and G. Malsiner-Walli. From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, 13(1):33–64, 2019. doi: 10.1007/s11634-018-0329-y.

S. Frühwirth-Schnatter, C. Pamminger, A. Weber, and R. Winter-Ebmer. Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *Journal of Applied Econometrics*, 27:1116–1137, 11 2012. doi: 10.1002/jae.1249.

S. Frühwirth-Schnatter, S. Pittner, A. Weber, and R. Winter-Ebmer. Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering. *The Annals of Applied Statistics*, 12:1796–1830, 09 2018. doi: 10.1214/17-AOAS1132.

S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, 1992.

A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2020. URL `https://CRAN.R-project.org/package=mvtnorm`. R package version 1.1-1.

B. Grün and F. Leisch. FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4):1–35, 2008. doi: 10.18637/jss.v028.i04.

J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. doi: 10.2307/2346830.

J. Hartzel, A. Agresti, and B. Caffo. Multinomial logit random effects models. *Statistical Modelling*, 1:81–102, 2001. doi: 10.1177/1471082x0100100201.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444.

D. Hedeker. *Multilevel Models for Ordinal and Nominal Variables*, chapter 6, pages 237–274. Springer-Verlag, 2008. ISBN 978-0-387-73183-4. doi: 10.1007/978-0-387-73186-5.

J. Jiang. *Linear and Generalized Linear Mixed Models and Their Applications.* Springer-Verlag, 1st edition, 2007. ISBN 978-0-387-47941-5. doi: 10.1007/978-0-387-47946-0.

A. Komárek and L. Komárková. Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*, 7(1):177–200, 2013. doi: 10.1214/12-aoas580.

A. Komárek and L. Komárková. Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software*, 59(12):1–38, 2014. doi: 10.18637/jss.v059.i12.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982. ISSN 0006341X, 15410420. doi: 10.2307/2529876.

J. S. Long. *Regression Models for Categorical and Limited Dependent Variables.* Sage, Thousand Oaks, CA, 1997.

D. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS – a Bayesian modeling framework: Concepts, structure and extensibility. *Statistics and Computing*, 10:325–337, 10 2000. doi: 10.1023/A:1008929526011.

G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün. Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26:303–324, 2016. doi: 10.1007/s11222-014-9500-2.

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.

G. Molenberghs and G. Verbeke. *Models for Discrete Longitudinal Data.* Springer, New York, 2005. ISBN 0-387-25144-8.

R. B. Nelsen. *An Introduction to Copulas*, volume 139 of *Lecture Notes in Statistics.* Springer, New York, 1999.

B. Peterson and F. E. Harrell. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(2):205–217, 1990.

J. C. Pinheiro and E. C. Chao. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):58–81, 2006. ISSN 10618600. doi: 10.1198/106186006x96962.

M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, 2003.

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006. URL https://journal.r-project.org/archive/.

C. Proust-Lima, V. Philipps, A. Diakite, and B. Liquet. Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software*, 78(2):1–56, 2017.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Second edition. Springer-Verlag, New York, 2004. ISBN 0-387-21239-6.

Stan Development Team. RStan: the R interface to Stan, 2020. URL http://mc-stan.org/. R package version 2.21.2.

M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

R. Stiratelli, N. M. Laird, and J. H. Ware. Random-effects models for serial observations with binary response. *Biometrics*, 40(4):961–971, 1984.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398): 528–550, 1987. doi: 10.2307/2289457.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000. ISBN 978-0-387-98784-2.

J. Vávra and A. Komárek. Classification based on multivariate mixed type longitudinal data: With an application to the EU-SILC database. *Advances in Data Analysis and Classification*, 2022. doi: https://doi.org/10.1007/s11634-022-00504-8.

J. Vávra, A. Komárek, B. Grün, and G. Malsiner-Walli. Clusterwise multivariate regression of mixed-type panel data. *Submitted, available as preprint*. doi: https://doi.org/10.21203/rs.3.rs-1882841/v1.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

G. Verbeke and E. Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221, 1996. doi: 10.1080/01621459.1996.10476679.

L. Villarroel, G. Marshall, and A. E. Barón. Cluster analysis using multivariate mixed effects models. *Statistics in Medicine*, 28(20):2552–2565, 2009. doi: 10.1002/sim.3632.

List of Figures

List of Tables

List of Algorithms