**Advances in Data Analysis and Classification Manuscript No.**
(to be inserted by the editor)

# Classification Based on Multivariate Mixed Type Longitudinal Data

## with an application to the EU-SILC database

**Jan Vávra · Arnošt Komárek**

Received: July 1, 2021 / Accepted: April 24, 2022

**Abstract** Although many present day studies gather data of a diverse nature (numeric quantities, binary indicators or ordered categories) on the same units repeatedly over time, there only exist limited number of approaches in the literature to analyse so-called *mixed-type* longitudinal data. We present a statistical model capable of joint modelling several mixed-type outcomes, which also accounts for possible dependencies among the investigated outcomes. A thresholding approach to link binary or ordinal variables to their latent numeric counterparts allows us to jointly model all, including latent, numeric outcomes using a multivariate version of the linear mixed-effects model. We avoid the independence assumption over outcomes by relaxing the variance matrix of random effects to a completely general positive definite matrix. Moreover, we follow model-based clustering methodology to create a mixture of such models to model heterogeneity in the temporal evolution of the considered outcomes. The estimation of such an hierarchical model is approached by Bayesian principles with the use of Markov chain Monte Carlo methods. After a successful simulation study with the aim to examine the ability to consistently estimate the true parameter values and thus discover the different patterns, the EU-SILC dataset consisting of Czech households that were followed for four years in a time span from 2005 – 2016 was analysed. The households were classified into groups with a similar evolution of several closely related indicators of monetary poverty based on estimated classification probabilities.

**Keywords** Multivariate longitudinal data · Mixed type outcome · Model based clustering · Classification · EU-SILC

## 1 Introduction

In different types of studies, data are nowadays routinely gathered repeatedly over time on the same units leading to *longitudinal* or *panel* data. In addition, multiple outcomes, both *numeric* and *categorical*, i.e., of a *mixed type*, are recorded at each measurement occasion leading to *multivariate mixed type longitudinal data*. An example of such a dataset, which also motivates our research, is *The European Union Statistics on Income and Living Conditions database* (EU-SILC, `https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions`). This is an instrument with the goal to collect timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions in the European Union, Iceland, Norway and Switzerland. The reference population includes all private households of the respective countries and variables, which are collected annually via questionnaires, and refer both to households and to individuals from the household. In this paper, we focus on household specific data from the Czech Republic (period 2005 – 2016) where each household was followed annually for a period of 4 years. In

Jan Vávra
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
Tel.: +420 951 553 385
E-mail: vavraj@karlin.mff.cuni.cz

Arnošt Komárek
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
Tel.: +420 951 553 282
E-mail: komarek@karlin.mff.cuni.cz

total, $n = 20\,323$ households will be analysed. The aim of the research is to find typical patterns of the temporal evolution of several indicators related to poverty and material deprivation. The relevant outcome variables are not only *numeric* (e.g., income) but also *binary* (e.g., ability to pay for a one week holiday), or *ordinal* (e.g., level of the financial burden of housing). Figure 1 illustrates such a combination of longitudinal outcomes used later (Section 6.1) in the analysis. From a data analytic point of view, it is our aim to develop a clustering approach suitable for longitudinal data of a mixed type which allows for the above-mentioned types of outcome variables.

To formalize the task, we are assuming that data are composed of $n$ independently behaving units (e.g., households) and for the $i$th unit ($i = 1,\ldots,n$), in total $R$ outcome variables $Y^r_{i,j}$ ($r = 1,\ldots,R$, $j = 1,\ldots,n_i$) are gathered at each of the $n_i$ measurement occasions that take place at times $t_{i,1},\ldots,t_{i,n_i}$. In addition, each outcome variable $Y^r_{i,j}$ might be either *numeric*, *binary* or *ordinal*. Finally, each observation might be supplemented by a vector $\boldsymbol{v}^r_{i,j}$ of additional covariates that may explain the outcome variability. In summary, the $i$th unit is represented by data $\mathscr{D}_i = \big\{Y^r_{i,j}, \boldsymbol{v}^r_{i,j}, t_{i,j} : r = 1,\ldots,R,\ j = 1,\ldots,n_i\big\}$, $i = 1,\ldots,n$ and the task is to use this information to classify each unit into one of $K > 1$ groups with a priori unknown structure.

Due to the complexity of a data structure and that possibly different numbers $n_i$ of measurement occasions appear in data for different units, classical distance-based clustering methods such as hierarchical clustering or the $K$-means method and their many extensions (see, e.g., Hastie et al, 2009, Chapters 13 and 14) could hardly be used. On the other hand, methods that further develop ideas of model based clustering (MBC, Banfield and Raftery, 1993; Fraley and Raftery, 2002) and that exploit mixtures of suitable statistical models proved to be useful in similar situations. Frühwirth-Schnatter (2006, Chapter 7) or more recently Grün (2019) provide a review of the Markov chain Monte Carlo (MCMC) methods for MBC. A classical model to analyse *continuous* longitudinal outcomes is the linear mixed model (LMM, Laird and Ware, 1982) and hence not surprisingly, several MBC procedures based on mixtures of LMM's appeared in the literature. The work by Verbeke and Lesaffre (1996), where growth curves are classified, provides one of the first methods of this type even though not explicitly called 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 ($R > 1$) outcome being available for the clustering, nevertheless, all of them still *continuous*, is considered by Villarroel et al (2009).

The MBC methods developed for functional data and (continuous) stochastic processes could also be employed if we continue to deal with *continuous* and, moreover, univariate ($R = 1$) longitudinal data (e.g. James and Sugar,

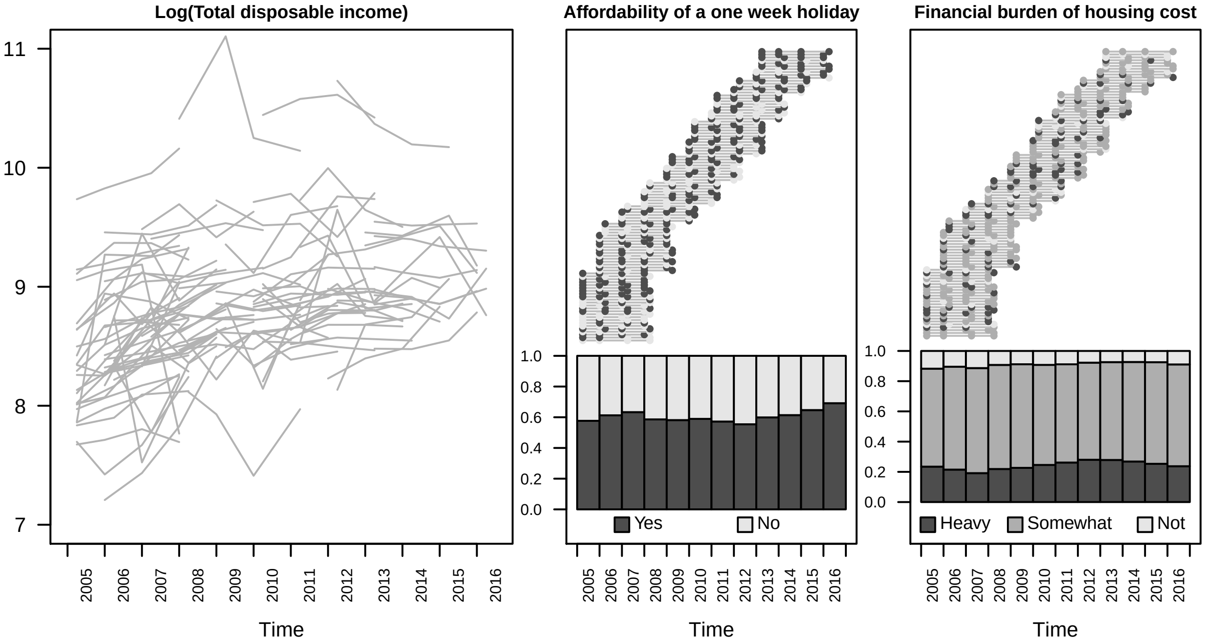


Fig. 1: EU-SILC data (Czech Republic). Observed longitudinal household profiles of three outcomes: Total disposable income (left, numeric, log-scale); Affordability of a one week holiday (centre, binary) and Financial burden of housing cost (right, ordinal). The shaded bars depict the proportions of the categorical outcome levels in each year.

2003; Ma et al, 2006; Liu and Yang, 2009; McNicholas and Murphy, 2010). Frühwirth-Schnatter (2011) provides a comprehensive overview. A possibility to develop the MBC for non-continuous longitudinal data is to replace LMM by a generalized linear mixed model (GLMM) in the underlying mixture of models. See, e.g., Molenberghs and Verbeke (2005, Chapter 14) who also provide an example of such a clustering procedure in their Section 23.3. Nevertheless, it is still only possible to use a single ($R = 1$) longitudinal outcome.

On the other hand, only little previous work appears to be available in the literature in cases where units are to be classified based on multivariate ($R > 1$) and possibly non-continuous longitudinal data. If all outcomes are of the same type (e.g., all binary), a method based again on a mixture of mixed models is offered by the R package lcmm (Proust-Lima et al, 2017). Nevertheless, for the MBC based on multivariate ($R > 1$) mixed type longitudinal data, the only two approaches we are aware of and that to some extent allow for classification, are those implemented in the R packages flexmix (Grün and Leisch, 2008) and mixAK (Komárek and Komárková, 2013, 2014). Nevertheless, both of the two approaches lack some important aspects. First, Grün and Leisch (2008) assume 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 $R$ outcomes in the clustering procedure. Even though a certain form of dependence is considered by Komárek and Komárková (2013, 2014), only binary or count non-continuous outcomes are considered, which does not allow for use with typical questionnaire data such as the EU-SILC database where many outcome variables are of an ordinal nature.

One of the reasons why there is not much available to perform clustering based on multivariate mixed type longitudinal data is perhaps the fact that even statistical models needed to develop the MBC procedure that would allow for datasets of a considered structure are relatively scarce in literature. This is especially if we seek models that realistically account for possible dependencies between different outcome variables gathered at one occasion. Fieuws and Verbeke (2004) covered in detail a bivariate case of longitudinal data and in this manuscript, we also follow their suggestion to use a multivariate mixed model while specifying a general covariance matrix for the joint distribution of all involved random effects. Later, Fieuws and Verbeke (2006) extended this approach to more than two outcomes by pairwise fitting and construction of pseudo-likelihood to avoid computational problems with a covariance matrix of a higher dimension. Nevertheless, MBC was not employed in any of those solutions. Recently, Bruckers et al (2016) invented a clustering algorithm that updates the pseudo-log-likelihood of the pairwise approach and reclassifies individuals until no change is made. This solution, however, lacks inclusion of the binary and ordinal outcomes that we aim to provide in this article.

The remainder of the paper is organized as follows. In Section 2, we first outline the approach capable of a joint modelling of mixed-type (numeric, binary and ordinal) longitudinal data. Second, in Section 3, we incorporate the developed model within the clustering procedure that allows usage of data with a structure analogous to that in Figure 1 and the classification of study units into groups with apriori unknown structure. Yet, Section 3 only provides a theoretical clustering concept, which assumes full knowledge of unknown parameters. The transition into a practically applicable procedure is provided in Section 4, which outlines details of a Bayesian approach towards this goal. Further, Section 5 evaluates clustering as well as the estimation capabilities of our approach on a simulation study. In Section 6, we apply our method to the EU-SILC database in order to discover clusters of different evolution patterns and for classifying each household. Finally, Section 7 summarizes the proposed methodology and discusses further possibilities on how to improve it in reaction to our findings from the applications.

## 2 Joint modelling of mixed-type longitudinal data

At each measurement occasion, $R$ outcomes (numeric, ordinal or binary) are observed on each study unit. Let $\mathscr{R} = \{1, \dots, R\} = \mathscr{R}^{\mathsf{Num}} \cup \mathscr{R}^{\mathsf{OB}}$, $\mathscr{R}^{\mathsf{OB}} = \mathscr{R}^{\mathsf{Ord}} \cup \mathscr{R}^{\mathsf{Bin}}$, denote the index set of observed outcomes that consists of indices of numeric outcomes ($\mathscr{R}^{\mathsf{Num}}$), ordinal outcomes ($\mathscr{R}^{\mathsf{Ord}}$) and binary outcomes ($\mathscr{R}^{\mathsf{Bin}}$). Let $\boldsymbol{Y}_i^r = \left(Y_{i,1}^r, \dots, Y_{i,n_i}^r\right)^\top$ be the vector of values of outcome $r \in \mathscr{R}$ of subject $i = 1, \dots, n$ observed at times $\boldsymbol{t}_i = (t_{i,1}, \dots, t_{i,n_i})$ together with additional covariates $\boldsymbol{v}_{i,1}^r, \dots, \boldsymbol{v}_{i,n_i}^r$. Further, let $\mathscr{C}_i^r = \left\{\boldsymbol{t}_i, \boldsymbol{v}_{i,1}^r, \dots, \boldsymbol{v}_{i,n_i}^r\right\}$ denote both the measurement times and the covariate values for the outcome $r$ of the $i$th subject. Finally, let

$$\mathbb{Y}_i = \left(\boldsymbol{Y}_i^r, r \in \mathscr{R}\right), \qquad \mathscr{C}_i = \left\{\mathscr{C}_i^r, r \in \mathscr{R}\right\} \tag{1}$$

denote all information (outcomes and covariate values) available for the $i$th subject, which is assumed to be independent of other subjects. $\boldsymbol{Y}^r$ and $\mathscr{C}^r$ stand for information (outcome and covariate values) regarding one chosen outcome $r \in \mathscr{R}$ from all subjects, while $\mathbb{Y}$ and $\mathscr{C}$ stand for all gathered information (all outcomes and covariate values) from all subjects.

The joint model for data (1) is built hierarchically. It exploits the linear mixed model (LMM) for each longitudinal outcome (each $r \in \mathscr{R}$). In the case of binary or ordinal outcomes, the LMM is assumed only latently. Dependencies between different outcomes gathered on a single study unit are captured by considering a vector of shared random effects. In particular, the model is built as follows.

## 2.1 Numeric longitudinal outcomes

For each numeric outcome $r \in \mathscr{R}^{\mathsf{Num}}$ we directly assume the linear mixed model:

$$\boldsymbol{Y}_i^r \mid \boldsymbol{b}_i^r; \mathscr{C}_i^r \sim \mathsf{N}_{n_i}\left(\boldsymbol{\eta}_i^r, \tau_r^{-1}\mathbb{I}_{n_i}\right), \tag{2}$$

where $\boldsymbol{\eta}_i^r = \mathbb{X}_i^r \boldsymbol{\beta}^r + \mathbb{Z}_i^r \boldsymbol{b}_i^r$ is the linear predictor consisting of fixed and random effects parts, $\tau_r > 0$ is the precision (inverse variance) of model errors , $\boldsymbol{\beta}^r \in \mathbb{R}^{d_r^{\mathsf{F}}}$ are fixed effects and $\boldsymbol{b}_i^r \in \mathbb{R}^{d_r^{\mathsf{R}}}$ are random effects belonging to subject $i$. Further, $\mathbb{X}_i^r = \left(\boldsymbol{x}_{i,1}^r, \dots, \boldsymbol{x}_{i,n_i}^r\right)^\top$ and $\mathbb{Z}_i^r = \left(\boldsymbol{z}_{i,1}^r, \dots, \boldsymbol{z}_{i,n_i}^r\right)^\top$ are matrices of regressors being derived from the explanatory variables information $\mathscr{C}_i^r$. For identifiability purposes, matrices $\mathbb{X}_i^r$ and $\mathbb{Z}_i^r$ are assumed not to share the same columns, i.e. the created regressor falls exclusively either into the fixed effects part or into the random effects part of the model.

## 2.2 Ordinal and binary longitudinal outcomes

The $r$th binary or ordinal outcome $r \in \mathscr{R}^{\mathsf{OB}}$ is assumed to attain values $0, \dots, L^r - 1$ which are linked to a linear mixed model through the thresholding concept (see, e.g., Albert and Chib, 1993):

$$Y_{i,j}^r = l \iff \gamma_l^r < Y_{i,j}^{\star,r} \leq \gamma_{l+1}^r, \tag{3}$$

where $-\infty = \gamma_0^r < \gamma_1^r < \cdots < \gamma_{L^r}^r = \infty$ are (unknown) thresholds categorizing a latent (unobserved) numeric variables $Y_{i,j}^{\star,r}$. Let $\boldsymbol{\gamma}^r = \left(\gamma_1^r, \dots, \gamma_{L^r-1}^r\right)$ be the vector of thresholds. For identifiability purposes, we will fix one of the thresholds, e.g., the first one $\gamma_1^r$ while estimating the model. That is, in the case of a binary outcome, all threshold parameters are fixed.

Analogously to the case of numeric outcomes, latent numeric variables $Y_{i,j}^{\star,r}$ are assumed to follow the linear mixed model

$$\boldsymbol{Y}_i^{\star,r} \mid \boldsymbol{b}_i^r; \mathscr{C}_i^r \sim \mathsf{N}_{n_i}\left(\boldsymbol{\eta}_i^r, \mathbb{I}_{n_i}\right) \tag{4}$$

with analogous notation to that in (2). Nevertheless, this time, the precision parameter $\tau_r$ of model errors is fixed and equal to one for identifiability purposes.

## 2.3 Joint model

Let $\mathbb{Y}_i^{\mathsf{N}} = \left(\boldsymbol{Y}_i^r, r \in \mathscr{R}^{\mathsf{Num}}\right)$ denote a vector of all numeric outcomes of subject $i$. Further, let $\mathbb{Y}_i^{\star,\mathsf{OB}} = \left(\boldsymbol{Y}_i^{\star,r}, r \in \mathscr{R}^{\mathsf{OB}}\right)$ be a vector of latent numeric variables behind all ordinal and binary outcomes. The subvectors of $\mathbb{Y}_i^\star := \left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}}\right)$ are assumed to follow linear mixed models (2) and (4) with a set of fixed effects $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}^r, r \in \mathscr{R}\right\}$ and an overall vector of random effects $\boldsymbol{b}_i = \{\boldsymbol{b}_i^r, r \in \mathscr{R}\}$, thus, forming a multivariate linear mixed model (MV LME).

In the following, let $\boldsymbol{b}_i^{\mathsf{N}} = \left\{\boldsymbol{b}_i^r, r \in \mathscr{R}^{\mathsf{Num}}\right\}$ and $\boldsymbol{b}_i^{\mathsf{OB}} = \left\{\boldsymbol{b}_i^r, r \in \mathscr{R}^{\mathsf{OB}}\right\}$ be random effects related to models for numeric and ordinal/binary longitudinal outcomes, respectively. The overall random effects vector $\boldsymbol{b}_i \equiv \left(\boldsymbol{b}_i^{\mathsf{N}}, \boldsymbol{b}_i^{\mathsf{OB}}\right)$ is now assumed to follow a multivariate normal distribution with a general covariance matrix, i.e., it is assumed

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_i^{\mathsf{N}} \\ \boldsymbol{b}_i^{\mathsf{OB}} \end{pmatrix} \overset{\text{iid}}{\sim} \mathsf{N}_{d^{\mathsf{R}}}\left(\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{\mathsf{N}} \\ \boldsymbol{\mu}^{\mathsf{OB}} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{\mathsf{N}} & \boldsymbol{\Sigma}^{\mathsf{NOB}} \\ \boldsymbol{\Sigma}^{\mathsf{OBN}} & \boldsymbol{\Sigma}^{\mathsf{OB}} \end{pmatrix}\right), \tag{5}$$

where $d^{\mathsf{R}} = d_{\mathsf{N}}^{\mathsf{R}} + d_{\mathsf{OB}}^{\mathsf{R}} = \sum_{r \in \mathscr{R}} d_r^{\mathsf{R}}$ is the total dimension of $\boldsymbol{b}_i$, $\boldsymbol{\mu} \in \mathbb{R}^{d^{\mathsf{R}}}$ is the (unknown) mean value of the random effects and $\boldsymbol{\Sigma} > 0$ is the unknown random effects covariance matrix. This matrix is left to be completely general, which captures possible dependencies between different longitudinal outcomes. Figure 2 demonstrates how the value of a correlation coefficient $\rho$ 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.

Throughout the manuscript, the notation $p(\cdot\,|\,\cdot)$ will stand for a conditional probability distribution function. Next to the fixed effects $\boldsymbol{\beta}$, mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the unknown parameters of the model are $\boldsymbol{\tau} := \left(\tau_r, r \in \mathscr{R}^{\mathsf{Num}}\right)$, precisions of the error terms of the LMM's for numeric outcomes and $\boldsymbol{\gamma} = \left\{\boldsymbol{\gamma}^r, r \in \mathscr{R}^{\mathsf{Ord}}\right\}$, thresholds for ordinal outcomes.

The outlined model implies the following likelihood based on the observed data:

$$\begin{aligned} p\left(\mathbb{Y}_i \middle| \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\tau}, \boldsymbol{\gamma}; \mathscr{C}_i\right) &= \int\!\!\int p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\mathsf{OB}}, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{b}_i \middle| \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\tau}, \boldsymbol{\gamma}; \mathscr{C}_i\right) \mathrm{d}\boldsymbol{b}_i\, \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}} = \\ &= \int\!\!\int \underbrace{p\left(\mathbb{Y}_i^{\mathsf{OB}} \middle| \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{\gamma}\right)}_{\text{thresholding (3)}} \cdot \underbrace{p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \middle| \boldsymbol{b}_i, \boldsymbol{\beta}, \boldsymbol{\tau}; \mathscr{C}_i\right)}_{\text{MV LME (2),(4)}} \cdot \underbrace{p\left(\boldsymbol{b}_i \middle| \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)}_{(5)} \mathrm{d}\boldsymbol{b}_i\, \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}. \end{aligned} \tag{6}$$

The probability density functions which are integrated in (6) are of the form

$$\begin{aligned} p\left(\mathbb{Y}_i^{\mathsf{OB}} \middle| \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{\gamma}\right) &= \prod_{r \in \mathscr{R}^{\mathsf{OB}}} \prod_{j=1}^{n_i} \left[\sum_{l=0}^{L^r-1} \mathbb{1}_{\{l\}}\left(y_{i,j}^r\right) \mathbb{1}_{(\gamma_l^r, \gamma_{l+1}^r]}\left(y_i^{\star,\mathsf{OB}}\right)\right], \\ p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \middle| \boldsymbol{b}_i, \boldsymbol{\beta}, \boldsymbol{\tau}; \mathscr{C}_i\right) &= \prod_{r \in \mathscr{R}^{\mathsf{Num}}} \prod_{j=1}^{n_i} \varphi\left(y_{i,j}^r; \eta_{i,j}^r, \tau_r^{-1}\right) \cdot \prod_{r \in \mathscr{R}^{\mathsf{OB}}} \prod_{j=1}^{n_i} \varphi\left(y_{i,j}^{\star,r}; \eta_{i,j}^r, 1\right), \\ p\left(\boldsymbol{b}_i \middle| \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) &= \varphi\left(\boldsymbol{b}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \end{aligned} \tag{7}$$

where $\varphi(\cdot; \boldsymbol{m}, \boldsymbol{S})$ is probability density function of multivariate normal distribution with mean $\boldsymbol{m}$ and variance matrix $\boldsymbol{S}$.

## 3 Model-based clustering framework

Classification of the subjects into one of $K$ latent subgroups with apriori unknown structure will be based on the model-based clustering procedure developed above the model introduced in Section 2 in which all parameters of the underlying linear mixed models might be group-specific. As it is usual in this context, let $U_i \in \{1, \ldots, K\}$ denote an unobservable group-allocation indicator for subject $i$ ($i = 1, \ldots, n$). We assume that the model for $i$-th subject if it belongs to the $k$-th group (given $U_i = k$, $k = 1, \ldots, K$) is described by the probability density function $p\left(\mathbb{Y}_i \,\middle|\, \boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\gamma}; \mathscr{C}_i\right)$ of the form (6), where $\left\{\boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\tau}^{(k)}\right\}$ is a set of (possibly) group-specific

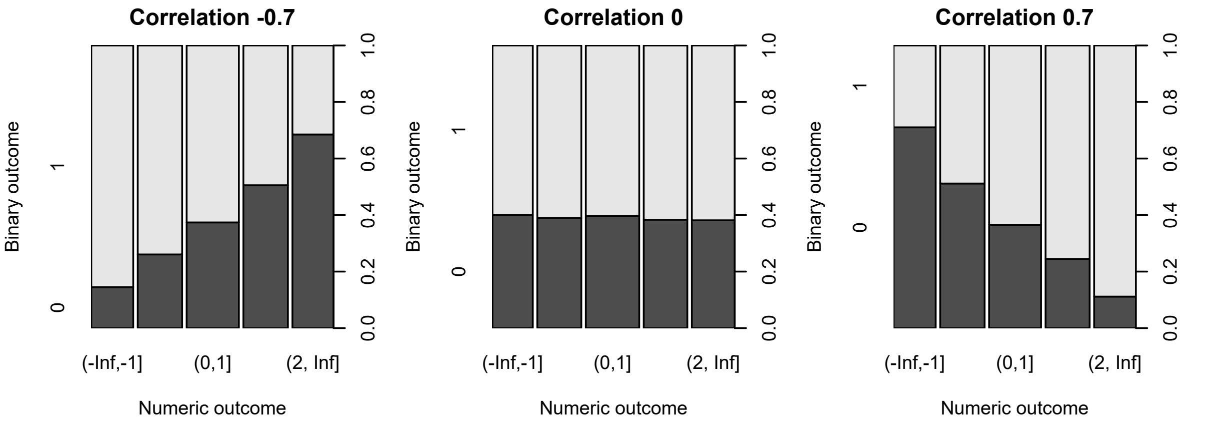


Fig. 2: Ratios of binary outcome values across different factorized values of the numeric outcome of the simulated longitudinal dataset for $n = 10\,000$ subjects each of $n_i = 4$ observations connected through random intercepts with correlation $\rho \in \{-0.7, 0, 0.7\}$.

model parameters. That is, the assumed conditional probability distribution function of the $i$-th subject outcomes given the group allocation is

$$p\left(\mathbb{Y}_i \,\middle|\, U_i = k, \boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\gamma}; \mathscr{C}_i\right) \overset{(6)}{=} \\ = \int\!\!\int p\left(\mathbb{Y}_i^{\mathsf{OB}} \,\middle|\, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{\gamma}\right) \cdot p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \,\middle|\, \boldsymbol{b}_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}; \mathscr{C}_i\right) \cdot p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right) \mathrm{d}\boldsymbol{b}_i \, \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}. \quad (8)$$

Note that by setting different LMM model parameters to be group-specific, we allow for different expressions of heterogeneity in the population. If, for example, we set parameters $\boldsymbol{\beta}$ to be group-specific, we assume that differences among the $K$ latent groups can be described in terms of the effect of the fixed effects covariates $\mathbb{X}_i$. On the other hand, group-specific parameter $\boldsymbol{\Sigma}$ would lead to different associations among random effects that would subsequently change the marginal relationships among the outcomes. In general, not all of the LMM model parameters must be group-specific; nevertheless, for clarity, we suppress this in notation. In the following, symbols $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\tau}$ will represent sets of all corresponding parameters $\left\{\boldsymbol{\beta}^{(k)}, k = 1, \ldots, K\right\}$, $\left\{\boldsymbol{\mu}^{(k)}, k = 1, \ldots, K\right\}$, $\left\{\boldsymbol{\Sigma}^{(k)}, k = 1, \ldots, K\right\}$ and $\left\{\boldsymbol{\tau}^{(k)}, k = 1, \ldots, K\right\}$, respectively.

Let $w_k = \mathsf{P}\big(U_i = k \,\big|\, \boldsymbol{w}\big) \in (0, 1)$, $k = 1, \ldots, K$, $\sum_{k=1}^{K} w_k = 1$, be the (unknown) probabilities of pertinence to each of the $K$ groups, $\boldsymbol{w} := \big(w_1, \ldots, w_K\big)$. Would we know all the model parameters $\boldsymbol{\theta} := \big\{\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\tau}, \boldsymbol{\gamma}\big\}$, Bayes rule provides an expression of conditional (given the observed data) probabilities for subject $i$ belonging to each of the groups:

$$u_{i,k}(\boldsymbol{\theta}) := \mathsf{P}\left[U_i = k \,|\, \mathbb{Y}_i, \boldsymbol{\theta}; \mathscr{C}_i\right] = \frac{w_k \, p\left(\mathbb{Y}_i \,\middle|\, U_i = k, \, \boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\gamma}; \, \mathscr{C}_i\right)}{\sum\limits_{k'=1}^{K} w_{k'} \, p\left(\mathbb{Y}_i \,\middle|\, U_i = k', \, \boldsymbol{\beta}^{(k')}, \boldsymbol{\mu}^{(k')}, \boldsymbol{\Sigma}^{(k')}, \boldsymbol{\tau}^{(k')}, \boldsymbol{\gamma}; \, \mathscr{C}_i\right)}. \quad (9)$$

In a majority of the MBC methodologies, the authors consider the maximum-likelihood estimation (MLE) of the unknown parameters, especially, its restricted version (REML) in the case of linear mixed models. The clustering is then based on estimated subject specific group probabilities $\widehat{u}_{i,k}^{\mathsf{ML}} = u_{i,k}\big(\widehat{\boldsymbol{\theta}}^{\mathsf{ML}}\big)$ where $\widehat{\boldsymbol{\theta}}^{\mathsf{ML}}$ denotes the MLE (or REML). In our situation, this likelihood takes the form of:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} w_k p\left(\mathbb{Y}_i \,\middle|\, U_i = k, \, \boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\gamma}; \, \mathscr{C}_i\right) \right\}.$$

This is traditionally solved by using the EM algorithm (Dempster et al, 1977) to face the problem of latent allocations leading to the mixture type likelihood. Nevertheless, two other levels of latent variables (random effects $\boldsymbol{b}_i$ and latent numeric variables $\mathbb{Y}_i^{\star,\mathsf{OB}}$) are present in our model leading to two additional levels of integration when evaluating the likelihood, see expression (6). This makes the likelihood hardly tractable and we switch to very popular Bayesian framework and the related MCMC methodology, which allows to fully exploit a hierarchical structure of our model. 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 maximization difficulties caused by subject-specific effects. The clustering itself is then based on the posterior distribution of the individual group probabilities (9) 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 pbc of a different type.

## 4 Bayesian inference

For Bayesian inference, we exploit the ideas of Bayesian data augmentation (BDA, Tanner and Wong, 1987) while considering all latent quantities, i.e., component allocations $\boldsymbol{U} := \big\{U_i, i = 1, \ldots, n\big\}$, LMM random effect vectors $\boldsymbol{b} := \big\{\boldsymbol{b}_i, i = 1, \ldots, n\big\}$ and latent variables $\mathbb{Y}^{\star,\mathsf{OB}} := \big\{\mathbb{Y}_i^{\star,\mathsf{OB}}, i = 1, \ldots, n\big\}$ as additional model parameters included in the posterior distribution. With the model specified in Sections 2 and 3, the joint distribution of observed as well as latent data and model parameters for the Bayesian model is given by the following decomposition

$$\begin{aligned}
p\big(\mathbb{Y}^{\mathsf{N}}, \mathbb{Y}^{\mathsf{OB}}, \mathbb{Y}^{\star,\mathsf{OB}}, \boldsymbol{U}, \boldsymbol{b}, \boldsymbol{\theta}; \mathscr{C}\big) &= \left[\prod_{i=1}^{n} p\big(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\mathsf{OB}}, \mathbb{Y}_i^{\star,\mathsf{OB}}, U_i, \boldsymbol{b}_i \mid \boldsymbol{\theta}; \mathscr{C}_i\big)\right] p(\boldsymbol{\theta}) \\
&= \left[\prod_{i=1}^{n} p\big(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\mathsf{OB}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \mid \boldsymbol{b}_i, U_i, \boldsymbol{\theta}; \mathscr{C}_i\big)\, p\big(\boldsymbol{b}_i \,\big|\, U_i, \boldsymbol{\theta}\big)\, p\big(U_i \,\big|\, \boldsymbol{\theta}\big)\right] p(\boldsymbol{\theta}) \\
&= \left[\prod_{i=1}^{n} p\big(\mathbb{Y}_i^{\mathsf{OB}} \,\big|\, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{\gamma}\big)\, p\big(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \,\big|\, \boldsymbol{b}_i, \boldsymbol{\beta}^{(U_i)}, \boldsymbol{\tau}^{(U_i)}; \mathscr{C}_i\big)\, p\big(\boldsymbol{b}_i \,\big|\, \boldsymbol{\mu}^{(U_i)}, \boldsymbol{\Sigma}^{(U_i)}\big)\, w_{U_i}\right] p(\boldsymbol{\theta}), \qquad (10)
\end{aligned}$$

where factors in (10) follow from (7) and $p(\boldsymbol{\theta})$ is the prior distribution of the primary model parameters. By symbols $\mathbb{Y}^{\mathsf{N}}$ and $\mathbb{Y}^{\mathsf{OB}}$ we understand the collection of corresponding outcomes of the same type across all individuals, i.e. $\mathbb{Y}^{\mathsf{N}} = \big\{\mathbb{Y}_i^{\mathsf{N}}, i = 1, \ldots, n\big\}$ and $\mathbb{Y}^{\mathsf{OB}} = \big\{\mathbb{Y}_i^{\mathsf{OB}}, i = 1, \ldots, n\big\}$. Later, we also use symbol $\mathbb{Y}$ for all outcomes available, that is $\mathbb{Y} = \mathbb{Y}^{\mathsf{N}} \cup \mathbb{Y}^{\mathsf{OB}}$.

### 4.1 Prior distribution and MCMC sampling scheme

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

$$p(\boldsymbol{\theta}) = p(\boldsymbol{w})\, p(\boldsymbol{\gamma})\, p\big(\boldsymbol{\beta} \,|\, \boldsymbol{\tau}\big)\, p(\boldsymbol{\tau})\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma})$$

with the following choices for the elements of the factorization.

A classically considered Dirichlet prior is assumed for the vector of group weights $\boldsymbol{w} = \big(w_1, \ldots, w_K\big)$, i.e.,

$$p(\boldsymbol{w}) \propto \prod_{k=1}^{K} w_k^{\alpha_k - 1},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is a set of positive hyperparameters (all being equal to 1 in our applications in Sections 5 and 6).

Considering the thresholding parameters $\boldsymbol{\gamma}^r$, $r \in \mathscr{R}^{\mathsf{Ord}}$, we first address the identifiability issue. Corresponding parametric space $\Omega^r$ is limited to a set of all vectors of ordered values with fixed first threshold $\gamma_1^r$. An improper uniform distribution on $\Omega^r$ is assumed for each set of thresholds $\boldsymbol{\gamma}^r$, $r \in \mathscr{R}^{\mathsf{Ord}}$. That is,

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

All fixed effects parameters $\boldsymbol{\beta}^{r,(k)} = \big(\beta_1^{r,(k)}, \ldots, \beta_{d_r^{\mathsf{F}}}^{r,(k)}\big)$, $r \in \mathscr{R}$, $k = 1, \ldots, K$, are assumed to be apriori independent and following a conjugate normal distributions, i.e.,

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{\tau}) = \prod_{k=1}^{K} \prod_{r \in \mathscr{R}^{\mathsf{Num}}} \prod_{j=1}^{d_r^{\mathsf{F}}} \varphi\left(\beta_j^{r,(k)}; \beta_{0,j}^{r}, \big(\tau_r^{(k)}\big)^{-1} d_{j,j}^{r}\right) \cdot \prod_{k=1}^{K} \prod_{r \in \mathscr{R}^{\mathsf{OB}}} \prod_{j=1}^{d_r^{\mathsf{F}}} \varphi\left(\beta_j^{r,(k)}; \beta_{0,j}^{r}, d_{j,j}^{r}\right),$$

where $\beta_{0,j}^{r}$ and $d_{j,j}^{r}$ are fixed hyperparameters (being equal to zero and ten, respectively, in our applications). The precision parameters are given independent gamma priors, i.e.,

$$p(\boldsymbol{\tau}) = \prod_{k=1}^{K} \prod_{r \in \mathscr{R}} p\big(\tau_r^{(k)}\big),$$

where each $p\big(\tau_r^{(k)}\big)$ corresponds to the gamma distribution $\Gamma(a_1, a_2)$. Also for the random effect means, a set of independent, and for simplicity, only semi-conjugate normal priors are assumed, i.e.,

$$p(\boldsymbol{\mu}) \equiv p(\boldsymbol{\mu} \,|\, \boldsymbol{\tau}_{\mathsf{R}}) = \prod_{k=1}^{K} \prod_{j=1}^{d^{\mathsf{R}}} p\big(\mu_j^{(k)} \,|\, \tau_{\mathsf{R},j}^{(k)}\big) = \prod_{k=1}^{K} \prod_{j=1}^{d^{\mathsf{R}}} \varphi\left(\mu_j^{(k)}; \mu_{0,j}^{(k)}, \big(\tau_{\mathsf{R},j}^{(k)}\big)^{-1}\right),$$

where $\mu_{0,j}^{(k)}$ are fixed hyperparameters (being equal to zero in our applications). Finally, parameters $\boldsymbol{\tau}_{\mathsf{R}} = \{\boldsymbol{\tau}_{\mathsf{R}}^{(k)}, k = 1,\dots,K\}$, $\boldsymbol{\tau}_{\mathsf{R}}^{(k)} = \left(\tau_{\mathsf{R},1}^{(k)}, \dots, \tau_{\mathsf{R},d^{\mathsf{R}}}^{(k)}\right)$ are random hyperparameters being assigned independent gamma priors $\Gamma(a_3, a_4)$ in another level of hierarchy to allow for weakly informative prior distribution. For calculations in Sections 5 and 6, $a_1 = a_3 = 1$ and $a_2 = a_4 = 1$ in the Gamma hyperpriors.

Covariance matrices $\boldsymbol{\Sigma}^{(k)}$ of random effects $\boldsymbol{b}_i$ are required to be completely general positive definite matrices, therefore, we suppose inverse covariance matrix $\boldsymbol{\Sigma}^{-(k)} := \left(\boldsymbol{\Sigma}^{(k)}\right)^{-1}$ to follow Wishart distribution to preserve conjugacy. Again, to achieve a weakly informative prior, we introduce a new hyperparameter, scale matrix $\mathbb{Q}^{(k)}$, while keeping the number of degrees of freedom $\nu_0 \geq d^{\mathsf{R}}$ fixed. Inverse $\mathbb{Q}^{-(k)}$ of auxiliary scale matrix $\mathbb{Q}^{(k)}$ is also assumed to follow Wishart distribution, this time with fixed diagonal scale matrix $\mathbb{D}^{\mathbb{Q}}$ and number of degrees of freedom $\nu_1$. In our applications, we use $\nu_0 = \nu_1 = d^{\mathsf{R}} + 1$ and $\mathbb{D}^{\mathbb{Q}} = 100 \cdot \mathbb{I}_{d^{\mathsf{R}}}$.

The related posterior distribution $p\big(\boldsymbol{\theta},\ \mathbb{Y}^{\star,\mathsf{OB}}, \boldsymbol{U}, \boldsymbol{b} \,\big|\, \mathbb{Y}^{\mathsf{N}}, \mathbb{Y}^{\mathsf{OB}};\ \mathscr{C}\big)$ and its characteristics are estimated using the MCMC methodology (Brooks et al, 2011). In particular, we adopted the classical Gibbs sampling scheme. Due to the (semi)-conjugate choices of prior distributions, all required full-conditioned distributions belong to some of the well-known distributional families and hence are straightforwardly sampled from them. See the Appendix for more details.

Finally, we point out that the posterior distribution is invariant towards the permutation of cluster labels. This may constitute a problem for classification. To avoid this *label switching* problem, we consider the post-sampling procedure of Stephens (2000) that considers all $K!$ permutations of labels for each iteration and ensures that the latent clusters $1,\dots,K$ have a fixed meaning during the whole sampling procedure. Only after label-switching has been addressed do we proceed with inference sensitive to the change of cluster labels, such as the estimation of classification probabilities.

## 4.2 Classification probabilities

Primarily, we perform classification using the posterior means

$$\widehat{U}_{i,k} = \int_{\boldsymbol{\theta}} u_{i,k}(\boldsymbol{\theta}) \cdot p\big(\boldsymbol{\theta} \,\big|\, \mathbb{Y}^{\mathsf{N}}, \mathbb{Y}^{\mathsf{OB}};\ \mathscr{C}\big)\, \mathrm{d}\boldsymbol{\theta}, \quad i = 1,\dots,n,\ k = 1,\dots,K \tag{11}$$

of allocation probabilities $u_{i,k}(\boldsymbol{\theta})$ defined in (9). With the MCMC based inference, we already have a sample (relabelled after label-switching check) from the posterior distribution $p\big(\boldsymbol{\theta} \,\big|\, \mathbb{Y}^{\mathsf{N}}, \mathbb{Y}^{\mathsf{OB}};\ \mathscr{C}\big)$ available and hence the values of (11) can be approximated by sample means of the respective values of $u_{i,k}(\boldsymbol{\theta})$ over the MCMC sample. Nevertheless, to evaluate the expression of $u_{i,k}(\boldsymbol{\theta})$ we need to calculate the probability density function (8), which involves non-trivial integrals with respect to the distribution of auxiliary latent variables: (i) the random effects $\boldsymbol{b}_i$ and (ii) the latent numeric outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$. In the rest of this section, we explain how to evaluate integrals in (8), that is, the integrals in

$$\begin{aligned} p\left(\mathbb{Y}_i \,\middle|\, U_i = k,\ \boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\gamma};\ \mathscr{C}_i\right) = \\ = \int p\left(\mathbb{Y}_i^{\mathsf{OB}} \,\middle|\, \mathbb{Y}_i^{\star,\mathsf{OB}},\ \boldsymbol{\gamma}\right) \cdot \underbrace{\left[\int p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \,\middle|\, \boldsymbol{b}_i,\ \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)};\ \mathscr{C}_i\right) \cdot p\big(\boldsymbol{b}_i \big| \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\big)\, \mathrm{d}\boldsymbol{b}_i\right]}_{p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \middle| \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)};\ \mathscr{C}_i\right)} \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}. \end{aligned} \tag{12}$$

### *4.2.1 Integration with respect to random effects* $\mathbf{b}_i$

Let us first integrate the random effects $\boldsymbol{b}_i$ out of (12) to obtain the marginal distribution of numeric variables. In this case, we will avoid integration by realizing that under the normality assumption of both numeric outcomes and random effects, the unconditioned distribution of the outcomes is also normal. Vector of all numeric and latent numeric outcomes $\boldsymbol{Y}_i$ ($\mathbb{Y}_i^{\mathsf{N}}$ combined with $\mathbb{Y}_i^{\star,\mathsf{OB}}$) of length $d = n_i|\mathscr{R}|$ given a vector of all random effects $\boldsymbol{b}_i$ follows by our LME assumption multivariate normal distribution:

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

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 $\mathbb{T}_i^{(k)}$ is diagonal due to independence assumption and contains the corresponding parameters of the residual variability, that is, $\left(\tau_r^{(k)}\right)^{-1}$ for $r \in \mathscr{R}^{\mathsf{Num}}$ and 1 otherwise. Using the normality of random effects, i.e., $\boldsymbol{b}_i \,|\, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)} \sim \mathsf{N}_{d^{\mathsf{R}}}\left(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right)$, and well known formulas for the moments, we obtain the marginal mean and the covariance matrix:

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

which results in

$$\boldsymbol{Y}_i \middle| \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)};\, \mathscr{C}_i \sim \mathsf{N}_d\left(\mathbb{X}_i\boldsymbol{\beta}^{(k)} + \mathbb{Z}_i\boldsymbol{\mu}^{(k)},\, \mathbb{V}_i^{(k)}\right). \tag{13}$$

This distribution has a general covariance structure, which also shows how our model captures dependencies among the longitudinal outcomes.

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

It remains to perform the following integration:

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

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. First, we separate marginal distribution of numeric outcomes $\mathbb{Y}_i^{\mathsf{N}}$ since it can avoid the integration, while the conditional normal distribution of latent numeric outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$ given $\mathbb{Y}_i^{\mathsf{N}}$ still awaits the integration:

$$\underbrace{p\left(\mathbb{Y}_i^{\mathsf{N}} \middle| \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)};\, \mathscr{C}_i\right)}_{\text{pdf of MVN}} \cdot \int \underbrace{p\left(\mathbb{Y}_i^{\mathsf{OB}} \middle| \mathbb{Y}_i^{\star,\mathsf{OB}},\, \boldsymbol{\gamma}\right)}_{\text{thresholding (3)}} \cdot \underbrace{\varphi\left(\mathbb{Y}_i^{\star,\mathsf{OB}}; \boldsymbol{\eta}_{\mathsf{OB}}^{(k)}, \mathbb{V}_{\mathsf{OB}}^{(k)}\right)}_{\text{pdf of } \mathbb{Y}_i^{\star,\mathsf{OB}} \middle| \mathbb{Y}_i^{\mathsf{N}}} \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}},$$

where $\boldsymbol{\eta}_{\mathsf{OB}}^{(k)}$ and $\mathbb{V}_{\mathsf{OB}}^{(k)}$ are the conditional mean and the covariance matrix of $\mathbb{Y}_i^{\star,\mathsf{OB}} \middle| \mathbb{Y}_i^{\mathsf{N}}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)};\, \mathscr{C}_i$.

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 $\boldsymbol{\eta}_{\mathsf{OB}}^{(k)}$ and the covariance matrix $\mathbb{V}_{\mathsf{OB}}^{(k)}$. For each individual categorical outcome $r \in \mathscr{R}^{\mathsf{OB}}$ and observation $j \in \{1, \ldots, n_i\}$ the value $y_{i,j}^r = l$ determines an interval given by the corresponding pair of $\gamma$ parameters, see (3):

$$Y_{i,j}^r = l \quad \Longrightarrow \quad Y_{i,j}^{\star,r} \in \left(\gamma_l^r,\, \gamma_{l+1}^r\right] =: \left(e_{ij}^r, f_{ij}^r\right].$$

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

$$I_k\left(\mathbb{Y}_i^{\mathsf{OB}}\right) = \int\limits_{\square\left(\boldsymbol{\gamma}, \mathbb{Y}_i^{\mathsf{OB}}\right)} p\left(\mathbb{Y}_i^{\star,\mathsf{OB}} \middle| \mathbb{Y}_i^{\mathsf{N}}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)};\, \mathscr{C}_i\right) \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}} = \int\limits_{\boldsymbol{e}_i}^{\boldsymbol{f}_i} \varphi\left(\boldsymbol{y};\, \boldsymbol{\eta}_{\mathsf{OB}}^{(k)}, \mathbb{V}_{\mathsf{OB}}^{(k)}\right) \mathrm{d}\boldsymbol{y}. \tag{14}$$

Finally, after the integrals $I_k$ for all $k = 1, \ldots, K$ are computed, the classification probabilities can be calculated proportionally:

$$u_{i,k}\left(\boldsymbol{\theta}\right) = \frac{w_k \cdot p\left(\mathbb{Y}_i^{\mathsf{N}} \,|\, \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)};\, \mathscr{C}_i\right) \cdot I_k\left(\mathbb{Y}_i^{\mathsf{OB}}\right)}{\sum\limits_{k'=1}^{K} w_{k'} \cdot p\left(\mathbb{Y}_i^{\mathsf{N}} \,|\, \boldsymbol{\beta}^{(k')}, \boldsymbol{\tau}^{(k')}, \boldsymbol{\mu}^{(k')}, \boldsymbol{\Sigma}^{(k')};\, \mathscr{C}_i\right) \cdot I_{k'}\left(\mathbb{Y}_i^{\mathsf{OB}}\right)}. \tag{15}$$

In order to compute integrals (14) needed in (15), we adopted an effective algorithm presented by Genz (1992), which is also based on the MCMC sampling. Since the approximation of such an integral is needed $K$-times for each generated state of the Gibbs sampling, the procedure is considerably time-consuming. The implemented function `pmvnorm` from the R package `mvtnorm` (Genz et al, 2019) is used in our applications.

## 4.3 Classification rules

Once we have estimated the posterior means of classification probabilities $\widehat{U}_{i,k}$ we perform the classification as follows. Naturally, for subject $i$ we choose cluster $k$ such that the corresponding estimated $\widehat{U}_{i,k}$ is the largest among the all $k = 1,\ldots,K$ values. However, that may not be the most fitting choice in cases where two of the clusters have both comparable and high probability.

In order to prevent misclassification, we may to allow subjects to remain unclassified when the decision is unclear. One way to accomplish this is that we classify a particular subject into the cluster with the highest probability only if it clearly overcomes the second largest probability. That is, when the difference between the two largest probabilities is higher than a chosen threshold. However, the choice of the value of this threshold for different values of $K$ would be another problem to be dealt with. On the other hand, with a Bayesian approach another tool to quantify uncertainty in classification is readily available. Next to the classification probabilities $\widehat{U}_{i,k}$ which are the posterior means of $u_{i,k}(\boldsymbol{\theta}) = \mathsf{P}\left[U_i = k \,|\, \mathbb{Y}_i,\, \boldsymbol{\theta};\, \mathscr{C}_i\right]$, we can additionally calculate the credible intervals for each $u_{i,k}(\boldsymbol{\theta})$. In our application, we make use of 95% highest posterior density (HPD) credible intervals and propose to classify subject $i$ into class $k$ with the highest classification probability $\widehat{U}_{i,k}$ if and only if its *lower* 95% HPD bound is still higher than any other *upper* 95% HPD bound of the remaining probabilities. Otherwise, subject $i$ remains unclassified. This procedure fills clusters with their most typical representatives and keep indecisive subjects aside. Unclassified subjects can then be additionally analysed to determine the pair (or potentially larger group) of clusters they are most associated with.

## 4.4 Number of groups

Throughout the paper, we treat the number of latent classes $K$ known and to be selected in advance. In most circumstances, however, there is no prior knowledge of the suitable value of $K$ to be used. Usual practice in this situation is to fit models with several values of $K$ and then to choose the one that optimizes one of the known criteria. To this end, we follow the steps of Aitkin et al (2009) and use the procedure based on exploration of the posterior distribution of deviances for models with different values of $K$.

Deviance is a general goodness-of-fit measure derived from the log-likelihood function. For given $K$, it is defined as

$$D^K(\boldsymbol{\theta};\mathbb{Y},\mathscr{C}) = -2\log p(\mathbb{Y}|\boldsymbol{\theta};\mathscr{C}) = -2\sum_{i=1}^{n}\log p(\mathbb{Y}_i|\boldsymbol{\theta};\mathscr{C}_i). \tag{16}$$

Aitkin et al (2009) propose to decide about the two values $K_1 < K_2$ of the number of groups on the basis of the posterior probability

$$\mathsf{P}\big[D^{K_1}(\boldsymbol{\theta};\mathbb{Y},\mathscr{C}) > D^{K_2}(\boldsymbol{\theta};\mathbb{Y},\mathscr{C}) \,\big|\, \mathbb{Y};\mathscr{C}\big] \tag{17}$$

that compares the deviances of the two nested models.

With the MCMC based inference, the quantity (17) is easily calculated as soon as the deviance value (16) is evaluated for each sampled value of the model parameters $\boldsymbol{\theta}$. In this respect, we note that the contribution of individual $i$ to the deviance is expressed as

$$-2\log p(\mathbb{Y}_i|\boldsymbol{\theta};\mathscr{C}_i) = -2\log\left[\sum_{k=1}^{K}\int\!\!\int p\left(\mathbb{Y}_i, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{b}_i, U_i = k \,\middle|\, \boldsymbol{\theta};\mathscr{C}_i\right)\,\mathrm{d}\boldsymbol{b}_i\,\mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}\right] \tag{18}$$

and includes the integration (12) for calculating classification probabilities, which has been described in Section 4.2. Using the same notation, we can write

$$D^K(\boldsymbol{\theta};\mathbb{Y},\mathscr{C}) = -2\sum_{i=1}^{n}\log\left[\sum_{k=1}^{K} w_k\cdot p\left(\mathbb{Y}_i^{\mathsf{N}}\,\middle|\,\boldsymbol{\beta}^{(k)},\boldsymbol{\tau}^{(k)},\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)};\mathscr{C}_i\right)\cdot I_k\left(\mathbb{Y}_i^{\mathsf{OB}}\right)\right], \tag{19}$$

where the denominator of (15) is inserted into the logarithm. Therefore, calculation of the deviance for one set of parameters $\boldsymbol{\theta}$ requires the calculation of the classification probabilities for every individual and hence is possibly extremely time-consuming.

## 5 Simulation

To demonstrate the functionality of our proposed approach, we performed a simulation study. To this end, data consisting of a numeric, a binary and an ordinal variable were generated while assuming different types of random effects structure. The only parameter distinguishing the latent groups ($K = 2$ or $K = 3$) 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.

### 5.1 Simulation design

Each type of response (numeric, binary and ordinal) is represented by only one longitudinally measured variable ($Y_{i,j}^{\mathsf{N}}$, $Y_{i,j}^{\mathsf{B}}$, $Y_{i,j}^{\mathsf{O}}$, $i = 1,\ldots,n$ and $j = 1,\ldots,n_i$). We set the number of observations per one subject $n_i$ to be fixed at $n_i = 4$ for each of the $n$ subjects, $n \in \{100, 500, 1000\}$, which also corresponds to the same amount of observations

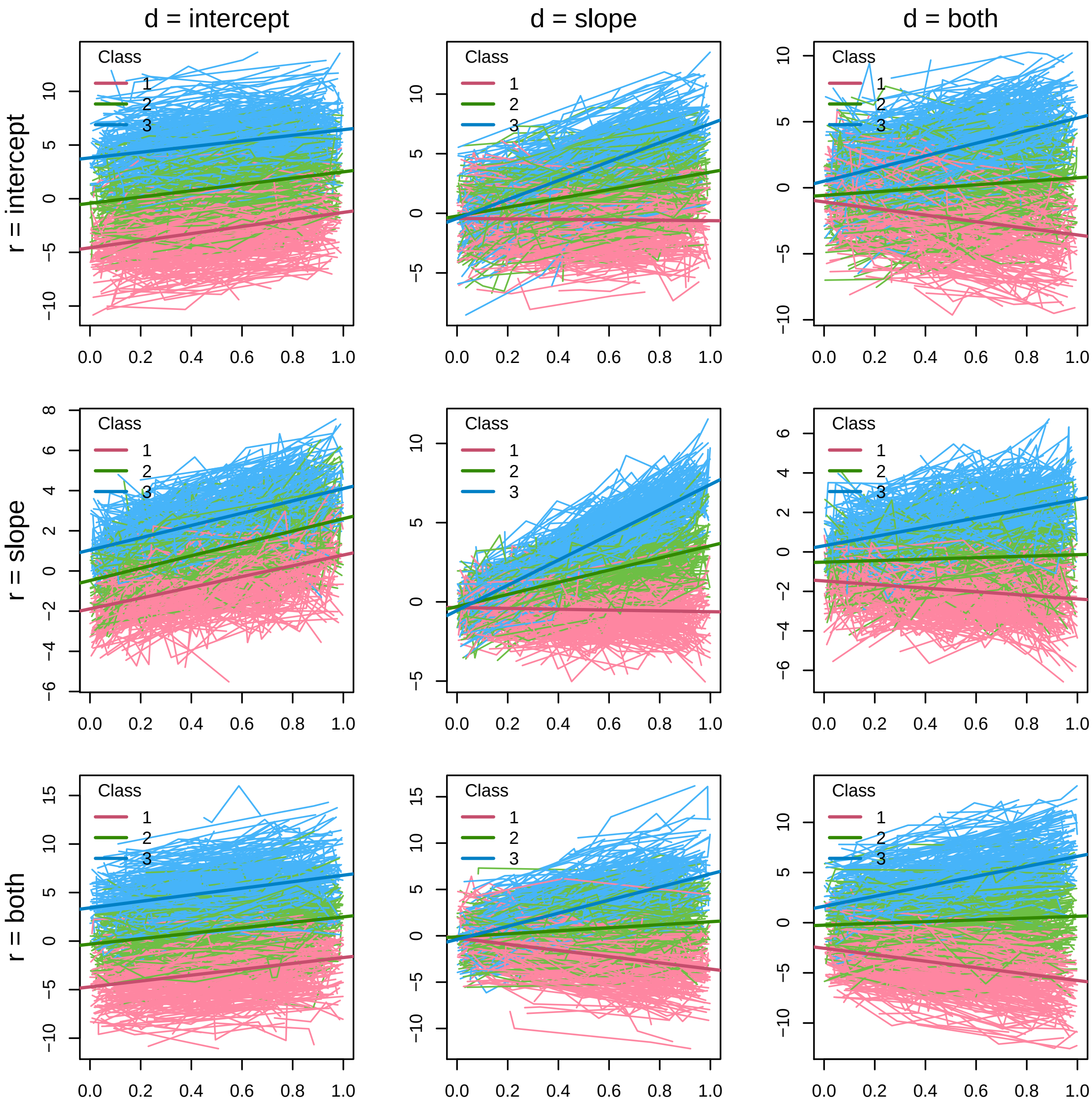


Fig. 3: The samples of a numeric outcome distinguishing different scenario types (row difference, column structure of random effects) when $K = 3$ latent groups are supposed.

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

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

Then, we suppose that each subject 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, which, however, depends on the structure of random effects. We consider three scenarios

1. (r = intercept): $b_{0,i} + \beta_1 t_{i,j}$, random intercept term and fixed slope,
2. (r = slope): $\beta_0 + b_{1,i}t_{i,j}$, fixed intercept term and random slope,
3. (r = both): $b_{0,i} + b_{1,i}t_{i,j}$, 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 subject are multivariate normal of dimension three (cases 1 and 2) or six (case 3). Its variance matrix $\boldsymbol{\Sigma}$ was an adequately chosen matrix with a non-diagonal form; more details can be found in supplementary materials. Another level of scenario settings arise from considering the three types of differences assumed among the $K = 2$ or $K = 3$ latent groups:

a) (d = intercept): only the intercept term $\beta_0^{(k)}$ (case 2) and $\boldsymbol{\mu}_0^{(k)}$ (case 1 and 3) are class-specific, but the slope parameters $\beta_1$ and $\boldsymbol{\mu}_1$ are not,

b) (d = slope): only the slope parameters $\beta_1^{(k)}$ (case 1) and $\boldsymbol{\mu}_1^{(k)}$ (case 2 and 3) are class-specific, but the intercept terms $\beta_0$ and $\mu_0$ are not,

c) (d = both): both the intercept and the slope terms $\beta_0^{(k)}$, $\beta_1^{(k)}$, $\boldsymbol{\mu}_0^{(k)}$ and $\boldsymbol{\mu}_1^{(k)}$ are class-specific.

These three types of differences are combined with the three types of random effects structure leading to nine different scenarios that are examined for $K = 2,3$ and different sample sizes $n$. 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 3 for the case $K = 3$). The true values of intercept and slope parameters can be found in Tables S2 and S3 in supplementary materials. 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 $\gamma_1^{\mathsf{B}} = 1$ and the ordinal variable by thresholds $\gamma_1^{\mathsf{O}} = -1$ and $\gamma_2^{\mathsf{O}} = 2$.

Each scenario under given $K$ 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 (14). 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 $n$ subjects would not take less than an hour even for the lowest values of $n = 100$ and $K = 2$ (around 80 minutes). The most challenging combination of $n = 1\,000$ and $K = 3$ took around 1 200 minutes. The number of calls of `pmvnorm` used for the approximation of posterior distribution of classification probabilities 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 $K = 3$).

### 5.2 Statistical properties

First, Figure 4 focuses on the properties of the posterior means of the model parameters being considered as classical estimators of the respective quantities. The colours distinguish the estimates in different classes ($k = 1,\ldots,K$) and the corresponding true values of the intercept and slope parameters are captured by dashed lines. The grey colour depicts the true value shared by all classes. Each segment represents 2.5% and 97.5% quantiles of 200 times replicated estimators and the full circle represents ite mean. Figure 4 provides estimates of parameters belonging to ordinal outcome only; plots for numeric and binary are postponed to supplementary materials.

Figure 4 demonstrates that the proposed procedure is capable of providing the estimators with reasonable statistical properties despite the latent modelling and the thresholding concept. In most cases, it successfully discovers the difference among classes 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 class-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 classes for low value of $n$ and the fact that LME usually tends to shrink random effects to zero. In the case of $K = 3$, this effect does not fully vanish even for $n = 1\,000$, see the row *both* and the column *intercept*. However, it seems that the large number of subjects $n$ can overcome this issue, which we rely on in the real data analysis shown in Section 6.

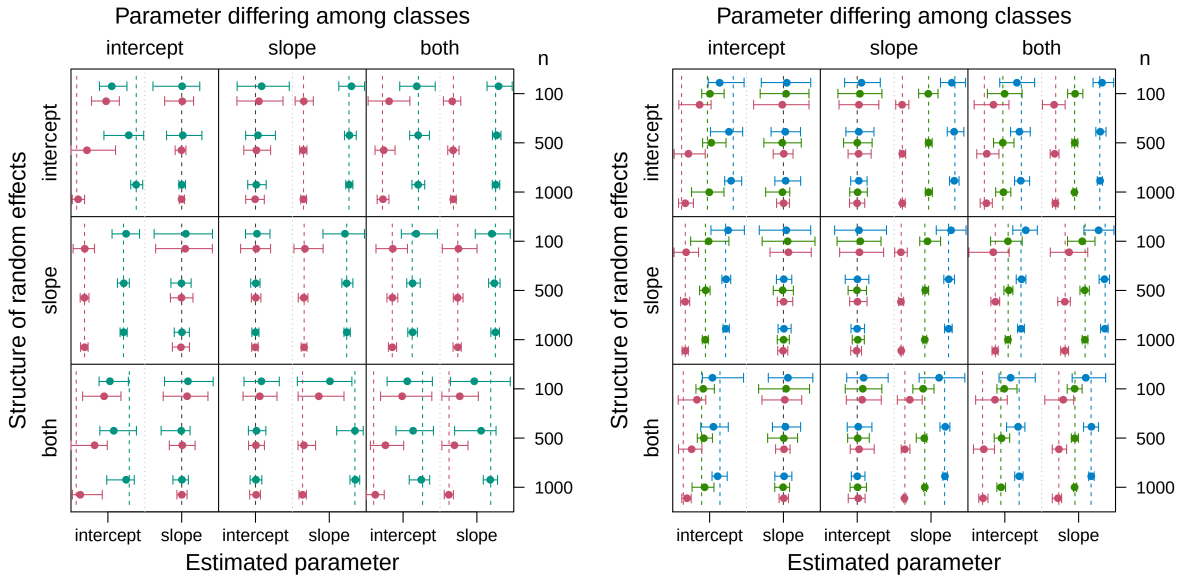
### 5.3 Classification ability

First, Table 1 contains the percentages of the correctly classified subjects (using the HPD interval rule) averaged across the 200 replications. This percentage differs scenario by scenario as the random structures and differences among the classes interact in different ways leading to diverse success rates. For example, the case with a class-specific random slope successfully classifies the vast majority of subjects for both $K = 2$ and $K = 3$, which is in agreement with the strict separation in the corresponding plot of Figure 3. Classification does not work satisfactorily in the problematic cases discussed above. Since for the low values of $n$, the difference between classes is not estimated to be as strict as it should be, a much larger percentage of subjects is kept unclassified in such cases. By increasing $n$, the percentage of unclassified subjects 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 $K = 3$ as one of the classes (green) is surrounded from both sides, which significantly reduces the ability to distinguish among classes, see Figure 3 for illustration.

The classification ability of our approach will also be evaluated by calculating the overall probability that a subject belonging to the $k$-th cluster is correctly classified in this cluster. To this end, for each $k$, we calculate the arithmetic mean $\overline{p}_k$ of the MCMC estimates $\widehat{U}_{i,k}$ of the posterior allocation probabilities (11) of belonging to cluster $k$ across all cluster members:

$$\overline{p}_k = \frac{1}{|i : U_i = k|} \sum_{i:U_i=k} \widehat{U}_{i,k}. \tag{20}$$

Further, to explore the impact of the longitudinally increasing amount of information, we also calculated the classification probabilities “dynamically”. This means that for each subject, we pretend a situation that subject $i$ is



(a) $K = 2$.

(b) $K = 3$.

Fig. 4: 95% quantile bounds and means for the intercept and slope parameters (separated by light grey dotted lines) for the ordinal outcome under different random effects structures and differences between classes. The true values of the parameters are depicted by dashed lines (dark grey if common to all classes).

Table 1: Percentages (standard deviation) of correctly classified, unclassified and misclassified subjects (using the HPD interval rule) for several choices of $n$, $K$, structure of random effects and class differences in 200 replications.

| r[1] | d[2] | $n$ | $K=2$ |  |  | $K=3$ |  |  |
|---|---|---|---|---|---|---|---|---|
|  |  |  | 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 classes.

to be classified on the basis of a set of first $j \in \{1,\ldots,n_i\}$ longitudinal observations that enter the expression (15) and consequently also the expression (20). Figure 5 shows the mean and the quantile bounds of such a dynamically calculated mean probabilities $\overline{p}_2$ based on 200 replications of experiments with $K = 3$ clusters. Class 2 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). The other choices of $k$ and $K$ (with much higher probabilities) can be found in the supplement.

If a difference among classes 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 classes, and therefore increasing the probability of classification towards the true middle class. It also improves with the increasing number of subjects $n$ since the classes are then better distinguished.

## 6 Application to EU-SILC data

We will now apply the proposed methodology in order to find the temporal patterns of the evolution of the chosen indicators for households in the EU-SILC database in the Czech Republic. The chosen time period of 2005 – 2016 covers the economical crisis and we expect it to heavily impact the budget of households, thus leading to different

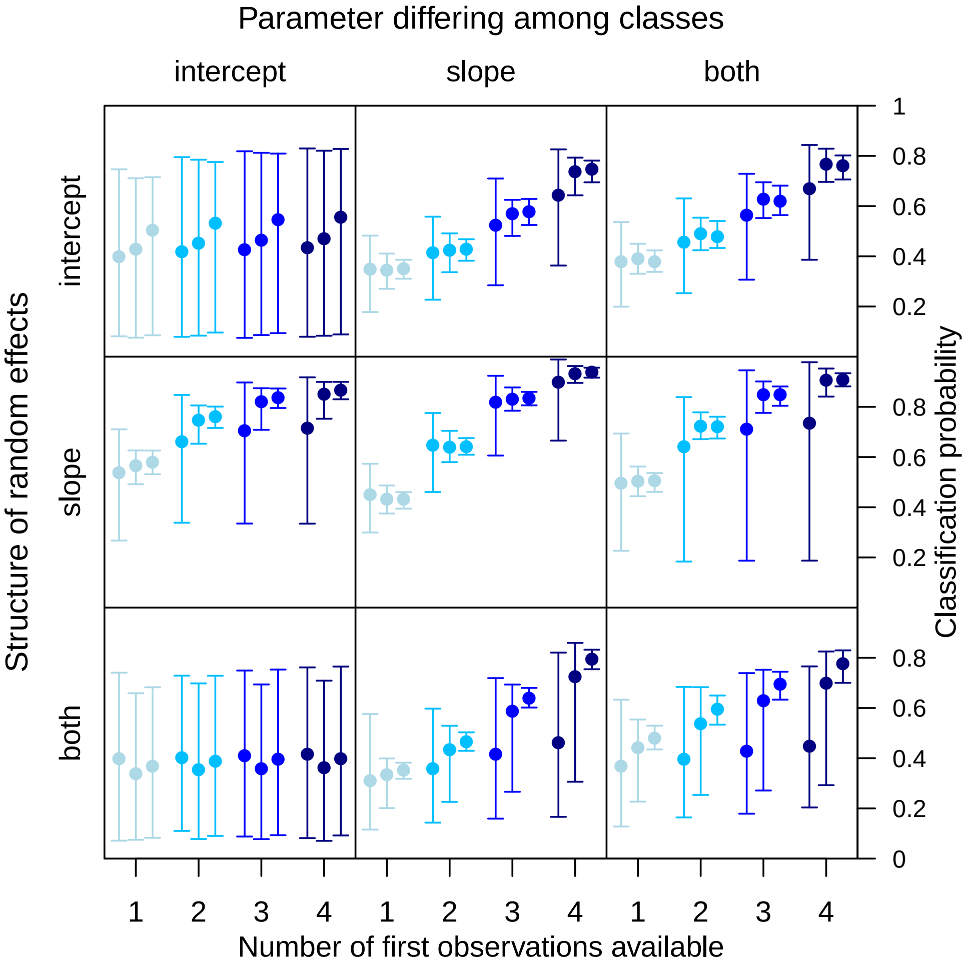


Fig. 5: Subjects of class 2 when $K = 3$. The mean and 2.5% and 97.5% quantile of mean classification probabilities $\overline{p}_2$ towards the true class calculated dynamically using only first $j \in \{1,2,3,4\}$ observations under several random effects structure and difference among $K = 3$ classes. Three lines of the same colour in one cell correspond to the increasing values of $n \in \{100,500,1000\}$.

ways of coping with the crisis. From those who were not affected and continue to prosper, to those who suffer unpleasant consequences. We chose one numeric, binary and ordinal variables that reflect the financial situation of a household the most and aim to discover several different patterns in their evolution while jointly modelling them.

## 6.1 Data description

First, we need to delve into the data gathering mechanism, which is crucial for the appropriate interpretation of the results. The EU-SILC longitudinal study follows a rotational design – rotating part of the sample from one year to the next and retaining the other unchanged part. The study in the Czech Republic was launched in 2005 with more than 7 000 households. Each following year, about a quarter of households in the study were dropped and replaced by newly entering ones. Apart from the natural exit from the study, households were followed for no longer than 4 years. Since the primary focus is on the evolution part, we use for the analysis only those households that were interviewed exactly $n_i = 4$ times. This decision reduces the number of total households used for the analysis to $n = 20\,323$.

The analysis will be performed on the following outcomes:

(i) *Total disposable income* (numeric),
(ii) *Capacity to afford paying for a one week annual holiday away from home* (binary - yes/no),
(iii) *Financial burden of the total housing cost* (ordinal - a heavy burden/a slight burden/no burden at all).

All-year income (in EUR) of the household (sum of all gross personal income components reduced by taxes on wealth, income and social insurance) follows heavily skewed distribution. Therefore, we work with its logarithmic transformation, which suits our LME assumptions much better. The binary outcome referred to as *Affordability of a one week holiday* has the actual meaning 'ability to pay' regardless of whether the household wants it. The ordinal outcome (in short *Financial burden of housing cost*) was filled subjectively by the respondent to assess his/her feeling about the extent to which housing costs are a financial burden to the household. For obvious reasons, these three cannot be considered as being completely independent.

The data contain information about the year and month of the interview (CZE data keep only the quarter – either Q1 or Q2). We construct the time variable as the number of years past the beginning of 2005, which limits the time into the interval $[0, 12)$. For the regression part of the model, we will also use the *Equivalised household size*, which is a sum of weights of each of the household members. The adult in the role of the head of the family has a weight of 1, others have either a weight of 0.5 or 0.3 depending on whether they are older or younger than 14, respectively.

## 6.2 Model structure

Clearly, the three outcomes are strongly related and we may benefit from modelling them jointly. We assume the LME model with the same structure of both fixed and random components for the numeric outcome and latent numeric counterparts of the binary and ordinal outcomes. Being aware of possible change in the evolution of these outcomes within the time period 2005 – 2016, we parametrize the effect of time by a B-spline of order 3 with knots in the years 2005, 2008, 2010 and 2017, which leads to six $\beta$ parameters including the intercept. This fixed part of the model is extended by the *Equivalised household size* as an additional regressor (denoted by $S$). The random effects structure, which is also responsible for the covariance structure among outcomes, is simply composed of the zero mean random intercept term, which allows households to evolve on a different level than others. The model formula for $j$-th observation of $i$-th household at time $t_{i,j}$ is then:

$$\underbrace{\beta_0^r + \beta_1^r B_1(t_{i,j}) + \cdots + \beta_5^r B_5(t_{i,j}) \ + \ \beta_6^r S_{i,j} +}_{\text{fixed effects}} \quad \underbrace{b_{0,i}^r}_{\text{random intercepts}} \quad , \quad r \in \{\mathsf{N}, \mathsf{B}, \mathsf{O}\},$$

where $B_1, \ldots, B_5$ are B-spline functions corresponding to spline of order 3 with knots at 0, 3, 5 and 12 that does not include the intercept and $\boldsymbol{b}_{0,i} = \left(b_{0,i}^{\mathsf{N}}, b_{0,i}^{\mathsf{B}}, b_{0,i}^{\mathsf{O}}\right)^\top$ is the three-dimensional mean-zero vector of random intercepts.

Our primary objective is to identify different patterns in the evolution of the chosen outcomes. Hence, fixed effects $\boldsymbol{\beta}^{(k)}$ are supposed to be cluster-specific leading to different patterns captured by the splines. Other model parameters ($\boldsymbol{\Sigma}$ and $\boldsymbol{\tau}$) responsible for the variance covariance structure are a nuisance with respect to our primary objective. To substantially reduce the dimension of parametric space and to concentrate on differences in patterns and not the dependency structure, we keep $\boldsymbol{\Sigma}$ and $\boldsymbol{\tau}$ the same in all classes.

### 6.3 Results

Contrary to the previous simulation study, we had to be more careful when sampling from the posterior distribution. The initial values of all the unknown parameters and latent variables were randomly generated to obtain different starting points for the sampled chains. The progress in each of the model parameters was visually monitored every 10 000th step in order to propose a reasonable choice of initial values for the subsequent continuation of sampling. Chains required up to hundred thousands iterations until the visual stationarity in all of the aspects was reached. The slow convergence was mainly caused by the threshold parameter $\gamma$ due to almost negligible steps. A final chain length of $M = 10\,000$ used for the analysis and results interpretation was sampled only after such visual stationarity was verified and then checked for label-switching issues. In the calculation of the classification probabilities and deviance we, again thinned the chain by 10 to save on computation time.

Following Section 4.4, we applied our methodology under several different choices of the number of hidden clusters $K$ and examined the posterior distribution of deviance in order to select the most suitable one. To this end, we searched for a value of $K$ where the decrease in deviance becomes negligible. Although, some improvement in the decrement of deviance is visible in Figure 6, we can also notice that the solution for $K = 4$ surprisingly achieved lower deviance than the one for $K = 5$. With $K = 5$ and also with higher values, small groups of households of extraordinary and very specific behaviour emerged leading us to the conclusion that from $K = 5$ onwards, we faced a clear overfitting problem, which can be also seen on Figure 7. This contains estimated splines curves for the logarithm of *Total disposable income* for each of the considered solutions. Case $K = 1$ shows us a general increasing trend flattening after the year 2011 (red curve), which seems to be followed by the majority of households even for higher $K$. With $K = 3$ a new violet cluster appears that follows the same shape of the general curve but on much higher level. It represents about 5–9% of households having a high income at their disposal. The solution for $K = 2$ actually started with parallel curves of the same shape, however, it slowly transformed one of the clusters into a very rare U-shaped trend (blue curve) cluster. For $K \geq 5$ such a cluster appears again accompanied by a golden cluster that behaves in reverse. This is why this solution should be rather viewed as an extension of the $K = 3$ solution. However, situation $K = 4$ avoids these clusters of extreme behaviour and additionally covers a turquoise cluster representing more than 10% of households of very low disposable income. This cluster seems to be the reason why this solution defeats $K = 5$ in terms of the deviance. The fact that the solution for $K = 5$ should rather be viewed as overfitting is also seen on the classification performance of the model. With $K = 5$, 25% of households remained unclassified), as the red and green clusters do not substantially differ. To interpret such clusters more precisely, we should not forget the other outcomes used. The resulting spline shapes for the

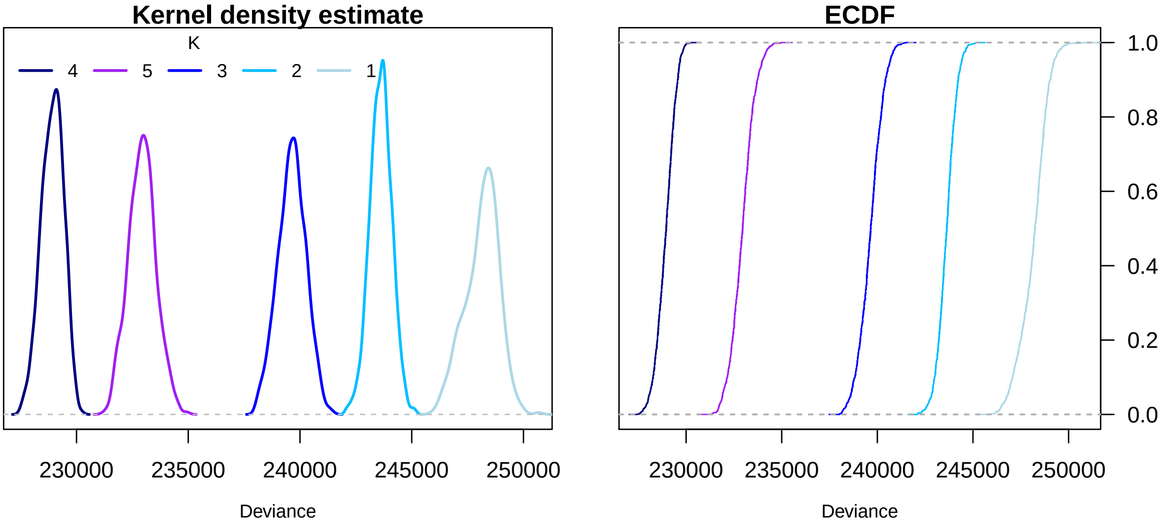


Fig. 6: Comparison of posterior distribution of deviances based on the model with the number of clusters $K = 1,2,3,4,5$.

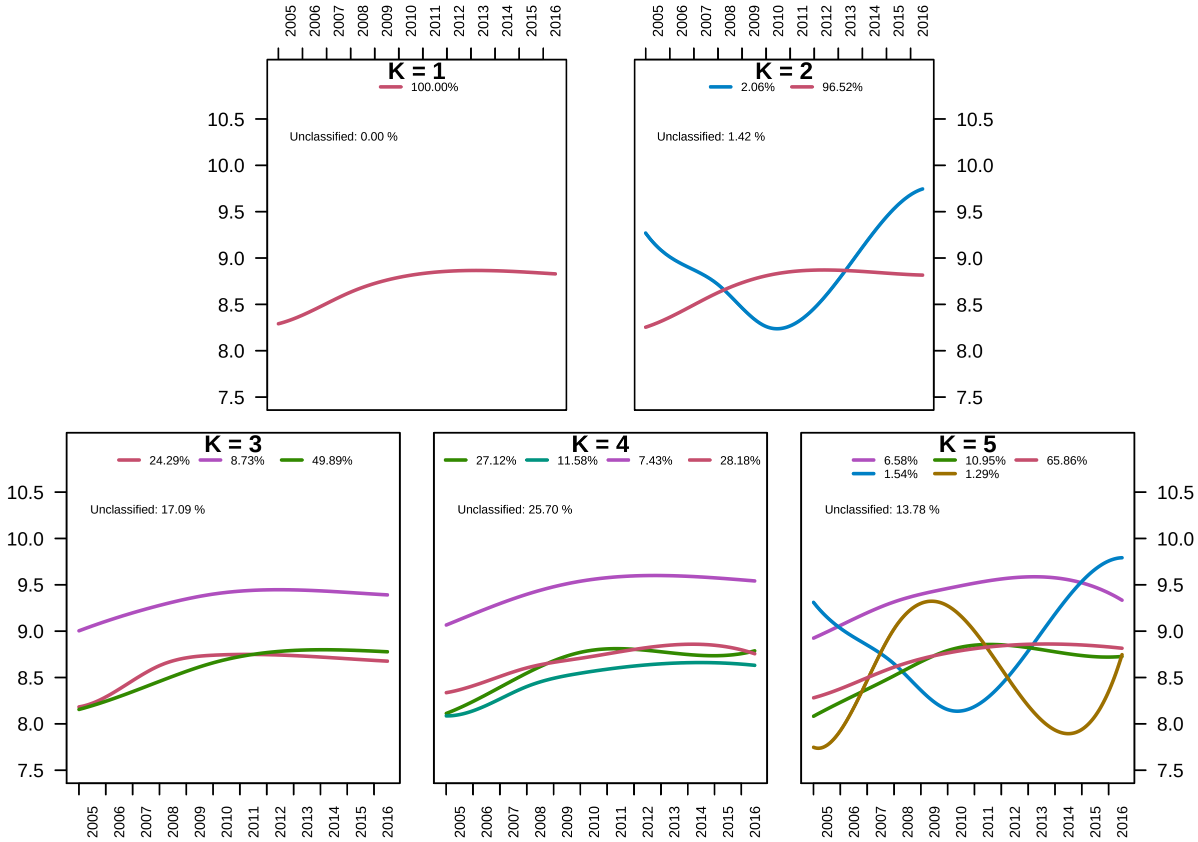


Fig. 7: Spline curves for the logarithm of the *Total disposable household income* of unit *Equivalised household size* for different choices of the number of clusters $K$.

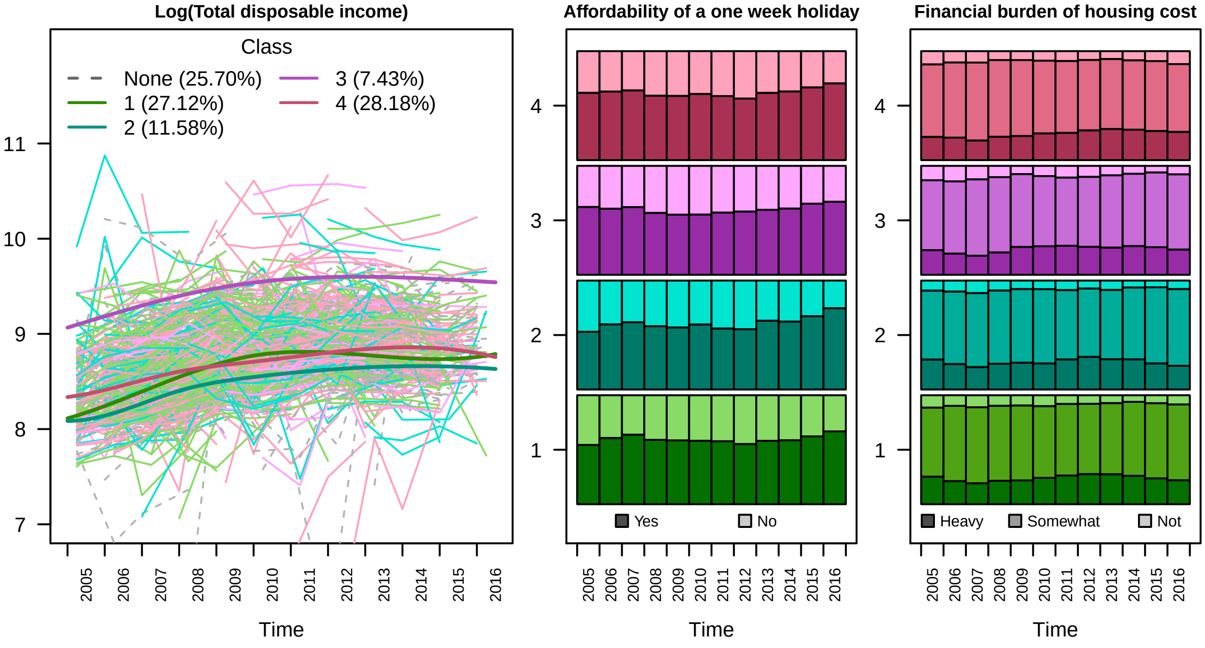


Fig. 8: Longitudinal profiles of numeric, binary and ordinal outcomes of $n = 1\,000$ randomly selected Czech households. Bold curves on the left represent the estimated conditional expectation of response within $K = 4$ discovered groups for a household of unit *Equivalised household size*. Categorical outcomes are shown by the proportions of categories in each year separately for the discovered groups. Some households remain unclassified.

*Affordability of a one week holiday* and the *Financial burden of total housing cost* can be found in the supplement, (Figures S14–S15).

Additional arguments on why the $K = 5$ solution is not satisfactory are as follows. As our goal is to find different patterns in evolution, we could have actually even been interested in the blue and golden clusters of extreme antagonistic behaviour. These clusters may cover households undergoing some substantial transformation, which may indeed be what we aim to identify. However, we must not forget the fact that households were only followed for 4 consecutive years. Therefore, the blue cluster should rather be interpreted in the following way: it consists of households measured in

- 2005 – 2009 - with rapidly decreasing income,
- 2009 – 2011 - with very low disposable income,
- 2011 – 2016 - with steeply increasing income,

but not necessarily following this trend for the whole span of 12 years, analogously for the golden cluster. Hence, these two clusters do not represent two of the typical outcome evolutions of a Czech household. This is the reason why we consider them rather an overfitting issue than actual clusters worthy of exploration. This leaves us with $K = 4$ solution being the most suitable for the overall interpretation.

In the $K = 4$ situation, households in the violet cluster with exceptionally high income can also always afford to pay for one week holiday abroad and do not find the housing cost to be a particularly heavy burden, see Figure 8. On the other hand, the turquoise cluster represents households of completely reverse characteristics - very low disposable income, inability to pay for a one week holiday abroad and almost all of them struggle with payments for housing. The other two remaining clusters (red and green) share a very similar and ordinary evolution of *Total disposable income*, but can still be distinguished. The proportions of categorical outcomes can be seen changing in time, especially the years 2010 and 2011 when the red cluster has the lowest percentage of households able to pay for a one week holiday, while the green cluster has the largest. Moreover, the evolution of proportions in both categorical outcomes is reversely mirrored; when one cluster thrives, the other struggles and vice versa. It almost seems like the large 60% cluster of average households was divided in half based on the undergoing positive or negative changes at certain periods of time. This division was enabled by our spline parametrization and the 4-year rotational panel invoked by the EU-SILC study.

## 7 Conclusion

First, we faced the problem of a joint modelling of longitudinally measured numeric, ordinal and binary outcomes. To this end, we proposed to use the multivariate linear mixed effects model on numeric and latent numeric outcomes corresponding to the categorical ones by exploiting the thresholding concept. While assuming all involved random effects follow a joint normal distribution, we enabled the longitudinal outcomes to be correlated. In addition, we considered the mixture of those models which allowed us to cluster individuals into several groups of various patterns in terms of the time evolution of the outcomes or the covariance structure. We allowed for specification of cluster-specific parameters, which provides the user with full flexibility to adjust their model to particular applications. The hierarchical nature of the model was then exploited within a fully Bayesian approach and the MCMC Gibbs sampling to estimate the model parameters and apply the clustering procedure based on the posterior distribution of the classification probabilities.

The proposed methodology was tested in a simulation study aimed at examining the ability to properly estimate model parameters and to correctly capture the patterns of each individual cluster. The results of the simulation study empirically show consistency of the parameter estimates, even in the case of categorical variables modelled by the thresholding approach. On the other hand, issues particularly related to the rate of convergence of the thresholds of latent numeric variables have been observed. Together with the slow approximation of the posterior distribution of classification probabilities caused by frequent integration of multivariate normal density over a multidimensional interval, it motivates us to replace the latent numeric outcome concept with suitable GLMMs (e.g. logit mixed-models), which are fully described by the model parameters and do not require any latent variables. Nevertheless, such models complicate the full-conditional distributions of unknown parameters, which requires sampling via the Metropolis proposal step instead of sampling directly from a well-known distributional family as in the current Gibbs procedure.

Another important issue, only marginally addressed by us is the choice of $K$, the total number of clusters. As an ultimate goal, nevertheless going far beyond the scope of this paper, we may attempt for an automated selection of the total number of clusters $K$. The use of a Dirichlet process mixture models by the lines of Neal (2000) might be considered. Frühwirth-Schnatter and Malsiner-Walli (2019) recently proposed a concept of sparse finite mixtures, which might also be considered in connection with our methodology.

Further, the covariance matrix $\boldsymbol{\Sigma}$ of random effects also needs 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 a general matrix $\boldsymbol{\Sigma}$ with a commonly used block structure.

So far, we have examined the properties of our methodology under a rather low number of outcomes. 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 part of the model. For example, in the EU-SILC database each household has several potential characteristics (family size, type of dwelling, number of rooms, degree of urbanization, region, country, gender, age or level of education of the head of a household, ...), where their influence on outcomes and subsequent clustering may be of interest.

Regarding the real data analysis, we successfully managed to discover several different patterns in the evolution of *Total disposable income*, *Affordability of a one week holiday* abroad and *Financial burden of the total housing cost*. Minorities of an extremely high (7.43%) or low (11.58%) living standard are easily distinguishable unlike the other mid-class households of similar income and categorical proportions but with different periods of increasing and decreasing tendencies. Using more than a four-cluster solution separates out households undergoing a huge progress or recession over a certain period of time. Although these findings may be of some interest, the corresponding patterns as a whole are unrealistic due to the rotational design of the study, which requires only observations from four consecutive years, and hence, these clusters are irrelevant from the realistic point of view. Maybe clustering not only based on the evolution over time itself but also on the covariance structure including relationships between the outcomes by allowing $\boldsymbol{\tau}$ and $\boldsymbol{\Sigma}$ to be cluster-specific could improve the cluster diversity.

Finally, the whole methodology was implemented as a set of C routines integrated into R (R Core Team, 2021). For those interested in applications, the implementation is provided via GitHub at `https://github.com/vavrajan/ClassNumOrdBin.git` together with a tutorial on how to use it. The R-C combination truly proved to be more efficient than a pure R implementation. The computation time, however, mainly depends on the number of subjects $n$ - the larger it is, the higher the number of latent parameters is to be sampled. For example, each of the cluster indicators $U_i$ requires computation of $K$ full conditional probabilities, which is still easily manageable as the latent numeric outcomes are at our disposal. This, however, does not hold for the classification probabilities (15) computed immediately after the sampling using an additional C function within R environment, since an integration over latent numeric outcomes need to be performed. The temporarily best solution (in terms of computation time and accuracy) involved calling a C version of `pmvnorm` function from the `mvtnorm` package which itself uses MCMC principles. Triggering this iterative process $K$-times for each individual and each sampled set of parameters takes a heavy toll.

**Acknowledgement**

This research was supported by the Czech Science Foundation (GAČR) grant 19-00015S. The first author was also supported by The Charles University Grant Agency, project GA UK No. 298120.

## Appendix

### A Full-conditioned distributions in Gibbs sampling

In this section, we will denote by $\boldsymbol{\Psi} = \{\boldsymbol{U}, \mathbb{Y}^{\star,\mathsf{OB}}, \boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{\tau}, \boldsymbol{\tau}_{\mathsf{R}}, \mathbb{Q}^{-1}\}$ the set of all parameters including randomized hyperparameters. We then derive full-conditioned distributions for all parameters $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ one by one. All derivations are based on viewing $p\big(\mathbb{Y} | \boldsymbol{\Psi};\ \boldsymbol{\zeta}, \mathscr{C}\big) \cdot p(\boldsymbol{\Psi} | \boldsymbol{\zeta})$ as a function of parameter $\boldsymbol{\psi}$, which can be decomposed into the following products:

$$\begin{aligned} p\left(\boldsymbol{\psi}\middle|\mathbb{Y},\boldsymbol{\Psi}_{-\boldsymbol{\psi}};\boldsymbol{\zeta},\mathscr{C}\right) \propto \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{OB}}\middle|\mathbb{Y}_i^{\star,\mathsf{OB}},\boldsymbol{\gamma}\right)\cdot \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{N}},\mathbb{Y}_i^{\star,\mathsf{OB}}\middle|\boldsymbol{b}_i,\boldsymbol{\beta}^{(U_i)},\boldsymbol{\tau}^{(U_i)};\mathscr{C}_i\right)\cdot \\ \cdot\prod_{i=1}^{n} p\left(\boldsymbol{b}_i\middle|\boldsymbol{\mu}^{(U_i)},\boldsymbol{\Sigma}^{-(U_i)}\right)\cdot\prod_{i=1}^{n} p\left(U_i|\boldsymbol{w}\right)\cdot \\ \cdot p\left(\boldsymbol{w}|\boldsymbol{\alpha}\right)\cdot p\left(\boldsymbol{\gamma}\middle|\gamma_1^r, r\in\mathscr{R}^{\mathsf{O}}\right)\cdot p\left(\boldsymbol{\beta}|\boldsymbol{\tau};\boldsymbol{\beta}_0,\mathbb{D}\right)\cdot p\left(\boldsymbol{\tau}|a_1,a_2\right)\cdot \\ \cdot p\left(\boldsymbol{\mu}|\boldsymbol{\tau}_{\mathsf{R}},\boldsymbol{\mu}_0\right)\cdot p\left(\boldsymbol{\tau}_{\mathsf{R}}|a_3,a_4\right)\cdot p\left(\boldsymbol{\Sigma}^{-1}\middle|\mathbb{Q}^{-1};\nu_0\right)\cdot p\left(\mathbb{Q}^{-1}\middle|\mathbb{D}^{\mathbb{Q}},\nu_1\right), \qquad (21) \end{aligned}$$

where $\boldsymbol{\zeta}$ denotes all fixed hyperparameters of prior distributions. Derivations are made under the assumption that parameters $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{\mu}$, $\boldsymbol{\tau}_{\mathsf{R}}$, $\boldsymbol{\Sigma}^{-1}$ and $\mathbb{Q}^{-1}$ are all group-specific. Similar derivations (with corresponding changes) can be made in the case of a chosen subset of group-specific parameters. Note that if $\boldsymbol{\tau}_{\mathsf{R}}$ and $\mathbb{Q}^{-1}$ are group-specific, then $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$ must also be group-specific.

### *A.1 Probabilities* $\boldsymbol{w}$

Prior probabilities $\boldsymbol{w}$ of belonging to a certain cluster, i.e. $w_k = \mathsf{P}\left(U_i = k\right)$, appear only in $p(U_i|\boldsymbol{w})$ and its prior distribution $p(\boldsymbol{w}|\boldsymbol{\alpha})$, therefore:

$$\begin{aligned} p\left(\boldsymbol{w}\,|\mathbb{Y},\boldsymbol{\Psi}_{-\boldsymbol{w}};\boldsymbol{\zeta},\mathscr{C}\right) &\propto \prod_{i=1}^{n} p\left(U_i|\boldsymbol{w}\right)\cdot p(\boldsymbol{w}|\boldsymbol{\alpha}), \\ p\left(\boldsymbol{w}\,|\boldsymbol{U};\boldsymbol{\alpha}\right) &\propto \prod_{k=1}^{K} w_k^{\sum_{i=1}^{n}\mathbb{1}(U_i=k)}\cdot\prod_{k=1}^{K} w_k^{\alpha_k-1} = \prod_{k=1}^{K} w_k^{n^k(\boldsymbol{U})+\alpha_k-1}, \end{aligned}$$

where $n^k(\boldsymbol{U})$ is the total number of appearances of value $k$ among all current group-allocation indicators $\boldsymbol{U} = \{U_i, i = 1,\dots,n\}$, i.e. the total number of subjects (from $n$ possible) currently belonging to group $k$. We recognize the shape of pdf of Dirichlet distribution, thus,

$$\boldsymbol{w}|\mathbb{Y},\boldsymbol{\Psi}_{-\boldsymbol{w}},\boldsymbol{\zeta};\mathscr{C} \sim \mathsf{Dir}_K\left(\boldsymbol{n}(\boldsymbol{U})+\boldsymbol{\alpha}\right),$$

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

### *A.2 Group-allocation indicators* $U_i$

According to (21), the group-allocation indicator $U_i$ appears only in its prior distribution $U_i|\boldsymbol{w}$ and at places, where it selects the corresponding group-specific parameter:

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

$U_i$ only attains values $k \in \{1,\dots,K\}$, therefore, we aim to calculate full-conditioned probability that $i$th subject is allocated in the group $k$:

$$\begin{aligned} &\mathsf{P}\left(U_i = k\middle|\mathbb{Y}_i^{\mathsf{N}},\mathbb{Y}_i^{\star,\mathsf{OB}},\boldsymbol{b}_i,\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1},\boldsymbol{w};\boldsymbol{\zeta},\mathscr{C}_i\right) \propto w_k\cdot\prod_{r\in\mathscr{R}^{\mathsf{Num}}}\left(\tau_r^{(k)}\right)^{\frac{n_i}{2}}\cdot\left|\boldsymbol{\Sigma}^{-(k)}\right|^{\frac{1}{2}}\cdot \\ &\cdot\exp\left\{-\frac{1}{2}\sum_{r\in\mathscr{R}^{\mathsf{Num}}}\sum_{j=1}^{n_i}\tau_r^{(k)}\left(y_{i,j}^r-\eta_{i,j}^{(k),r}\right)^2-\frac{1}{2}\sum_{r\in\mathscr{R}^{\mathsf{OB}}}\sum_{j=1}^{n_i}\left(y_{i,j}^{\star,r}-\eta_{i,j}^{(k),r}\right)^2-\frac{1}{2}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)^\top\boldsymbol{\Sigma}^{-(k)}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)\right\}, \end{aligned}$$

where $\eta_{i,j}^{(k),r} = \left(\boldsymbol{x}_{i,j}^r\right)^\top\boldsymbol{\beta}^{(k),r}+\left(\boldsymbol{z}_{i,j}^r\right)^\top\boldsymbol{b}_i^r$ is the linear predictor of $j$-th observation of outcome $r\in\mathscr{R}$ of $i$th subject when belonging to the group $k$.

### A.3 Latent numeric variables $\mathbb{Y}_i^{\star,\mathsf{OB}}$

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

$$p\left(\mathbb{Y}_i^{\star,\mathsf{OB}}\middle|\mathbb{Y},\boldsymbol{\Psi}_{-\mathbb{Y}_i^{\star,\mathsf{OB}}};\boldsymbol{\zeta},\mathscr{C}\right) \propto p\left(\mathbb{Y}_i^{\mathsf{OB}}\middle|\mathbb{Y}_i^{\star,\mathsf{OB}},\boldsymbol{\gamma}\right)\cdot p\left(\mathbb{Y}_i^{\mathsf{N}},\mathbb{Y}_i^{\star,\mathsf{OB}}\middle|\boldsymbol{b}_i,\boldsymbol{\beta}^{(U_i)},\boldsymbol{\tau}^{(U_i)};\mathscr{C}_i\right).$$

From (7), we see that for all $r \in \mathscr{R}^{\mathsf{OB}}$ and $j = 1,\dots,n_i$ are $Y_{i,j}^{\star,r}$ independently distributed. Ignoring the thresholding concept $Y_{i,j}^{\star,r}$ would follow $\mathsf{N}\left(\eta_{i,j}^{(U_i),r},1\right)$, however, corresponding density is now limited by indicator $\mathbb{1}_{(\gamma_l^r,\gamma_{l+1}^r]}\left(y_i^{\star,\mathsf{OB}}\right)$, where $l = y_{i,j}^r$. Therefore, the full-conditioned distribution is truncated normal distribution on the interval $\left(\gamma_l^r\,,\,\gamma_{l+1}^r\right]$:

$$Y_{i,j}^{\star,r}\Big|Y_{i,j}^r = l,\boldsymbol{\gamma} \sim \mathsf{TN}\left(\eta_{i,j}^{(U_i),r},1,\gamma_l^r,\gamma_{l+1}^r\right).$$

### A.4 Thresholds $\boldsymbol{\gamma}$

Parameter $\boldsymbol{\gamma}$ influences (21) only in the thresholding phase and in the prior distribution of $\boldsymbol{\gamma}$:

$$p\left(\boldsymbol{\gamma}\middle|\mathbb{Y},\boldsymbol{\Psi}_{-\boldsymbol{\gamma}^r};\boldsymbol{\zeta},\mathscr{C}\right) \propto \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{OB}}\middle|\mathbb{Y}_i^{\star,\mathsf{OB}},\boldsymbol{\gamma}\right)\cdot p\left(\boldsymbol{\gamma}\middle|\gamma_1^r, r\in\mathscr{R}^{\mathsf{O}}\right).$$

Let us consider ordinal outcome $r \in \mathscr{R}^{\mathsf{Ord}}$ and the corresponding set of thresholds: $-\infty = \gamma_0^r, \gamma_1^r, \boldsymbol{\gamma}^r, \gamma_{L^r} = \infty$. Let $\mathscr{Y}_l^r$ be the set of all latent numeric outcomes $Y_{i,j}^{\star,r}$ such that the truly measured ordinal category is $l = 0,\dots,L^r-1$, i.e.

$$\mathscr{Y}_l^r = \left\{Y_{i,j}^{\star,r} : Y_{i,j}^r = l, \quad i = 1,\dots,n, \quad j = 1,\dots,n_i\right\},$$

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

$$-\infty < \underset{\in\mathscr{Y}_0^r}{y_0} < \gamma_1^r < \underset{\in\mathscr{Y}_1^r}{y_1} < \gamma_2^r < \underset{\in\mathscr{Y}_2^r}{y_2} < \dots < \gamma_{L^r-1}^r < \underset{\in\mathscr{Y}_{L^r-1}^r}{y_{L^r-1}} < \infty.$$

Thus, under the uniform prior for $\boldsymbol{\gamma}^r$ (set in Section 4.1) we get that the individual thresholds $\gamma_l^r$ are uniformly distributed on intervals given by maxima and minima of the corresponding sets:

$$\gamma_l^r|\boldsymbol{Y}^r,\boldsymbol{Y}^{\star,r} \;\sim\; \mathsf{Unif}\left[\max_{y\in\mathscr{Y}_{l-1}^r} y, \min_{y\in\mathscr{Y}_l^r} y\right], \qquad l = 1,\dots,L^r.$$

### A.5 Precision parameters $\boldsymbol{\tau}$

Parameters $\boldsymbol{\tau} = \left\{\tau_r^{(k)} : k = 1,\dots,K,\; r \in \mathscr{R}^{\mathsf{Num}}\right\}$ are the inverse variance of errors of the supposed LME models over numeric outcomes. The right-hand side of (21) includes $\boldsymbol{\tau}$ only in three factors: the supposed LME for $\mathbb{Y}_i^{\mathsf{N}}$ and the prior distribution of $(\boldsymbol{\beta},\boldsymbol{\tau})$:

$$p\left(\boldsymbol{\tau}|\mathbb{Y},\boldsymbol{\Psi}_{-\boldsymbol{\tau}};\boldsymbol{\zeta},\mathscr{C}\right) \propto \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{N}},\mathbb{Y}_i^{\star,\mathsf{OB}}\middle|\boldsymbol{b}_i,\boldsymbol{\beta}^{(U_i)},\boldsymbol{\tau}^{(U_i)};\mathscr{C}_i\right)\cdot p\left(\boldsymbol{\beta}|\boldsymbol{\tau};\boldsymbol{\beta}_0,\mathbb{D}\right)\cdot p\left(\boldsymbol{\tau}|a_1,a_2\right).$$

From the structure of (7) and priors $p(\boldsymbol{\beta}|\boldsymbol{\tau})$ and $p(\boldsymbol{\tau})$ (set in Section 4.1) we see that individual $\tau_r^{(k)}$ are distributed independently of each other (given all other information and parameters):

$$\begin{aligned} p\left(\tau_r^{(k)}\middle|\boldsymbol{Y}^r,\boldsymbol{U},\boldsymbol{b}^r,\boldsymbol{\beta}^{(k),r};\boldsymbol{\beta}_0^r,\mathbb{D}^r,a_1,a_2,\mathscr{C}^r\right) &\propto \left(\tau_r^{(k)}\right)^{\frac{1}{2}\sum_{i=1}^{n} n_i\mathbb{1}(U_i=k)+\frac{1}{2}d_r^{\mathsf{F}}+a_1-1}\cdot \\ &\cdot\exp\left\{-\tau_r^{(k)}\left[\frac{1}{2}\sum_{i\in\mathscr{N}_k(\boldsymbol{U})}\sum_{j=1}^{n_i}\left(y_{i,j}^r-\eta_{i,j}^{(k),r}\right)^2+\frac{1}{2}\sum_{j=1}^{d_r^{\mathsf{F}}}\frac{\left(\beta_j^{(k),r}-\beta_{0,j}^r\right)^2}{d_{j,j}^r}+a_2\right]\right\}, \end{aligned}$$

where $\mathscr{N}_k(\boldsymbol{U}) = \{i : U_i = k,\ i = 1,\ldots,n\}$ is a set of subjects currently belonging to group $k$. For $\boldsymbol{Y}^r, \mathscr{C}^r$ and current values of $\boldsymbol{U}, \boldsymbol{b}^r$ and $\boldsymbol{\beta}^{(k)}$ let us denote

$$\widetilde{a}_1^{(k),r} = \frac{1}{2} \sum_{i \in \mathscr{N}_k(\boldsymbol{U})} n_i + \frac{d_r^{\mathsf{F}}}{2} + a_1,$$

$$\widetilde{a}_2^{(k),r} = \frac{1}{2} \sum_{i \in \mathscr{N}_k(\boldsymbol{U})} \sum_{j=1}^{n_i} \left( y_{i,j}^r - \eta_{i,j}^{(k),r} \right)^2 + \frac{1}{2} \sum_{j=1}^{d_r^{\mathsf{F}}} \frac{\left( \beta_j^{(k),r} - \beta_{0,j}^r \right)^2}{d_{j,j}^r} + a_2.$$

Under this notation we see that

$$\tau_r^{(k)} \left| \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{\beta}^{(k),r}; \boldsymbol{\beta}_0^r, \mathbb{D}^r, a_1, a_2, \mathscr{C}^r \right. \sim \Gamma\left( \widetilde{a}_1^{(k),r}, \widetilde{a}_2^{(k),r} \right)$$

independently for each $r \in \mathscr{R}^{\mathsf{Num}}$ and $k = 1,\ldots,K$.

### *A.6 Fixed effects* $\boldsymbol{\beta}$

Fixed effects $\boldsymbol{\beta}$ appear only in the LME model specification and prior distribution:

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

which can be decomposed for individual outcomes $r \in \mathscr{R}$ and $k = 1,\ldots,K$ as follows:

$$p\left( \boldsymbol{\beta}^{(k),r} \middle| \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \tau_r^{(k)}; \boldsymbol{\beta}_0^r, \mathbb{D}^r, \mathscr{C}^r \right) \propto \exp\left\{ -\frac{\tau_r^{(k)}}{2} \left( \boldsymbol{\beta}^{(k),r} - \boldsymbol{\beta}_0^r \right)^\top [\mathbb{D}^r]^{-1} \left( \boldsymbol{\beta}^{(k),r} - \boldsymbol{\beta}_0^r \right) \right\} \cdot$$
$$\cdot \exp\left\{ -\frac{\tau_r^{(k)}}{2} \left( \widetilde{\boldsymbol{y}}_{\mathscr{N}_k(\boldsymbol{U})}^r - \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r \boldsymbol{\beta}^{(k),r} \right)^\top \left( \widetilde{\boldsymbol{y}}_{\mathscr{N}_k(\boldsymbol{U})}^r - \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r \boldsymbol{\beta}^{(k),r} \right) \right\},$$

where notation $\bullet_{\mathscr{N}_k(\boldsymbol{U})}$ restricts given expression $\bullet$ to the subset of subjects in group $k$:

$$\mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r = \begin{pmatrix} \vdots \\ \mathbb{X}_i^r \\ \vdots \end{pmatrix}, i \in \mathscr{N}_k(\boldsymbol{U}), \qquad \widetilde{\boldsymbol{y}}_{\mathscr{N}_k(\boldsymbol{U})}^r = \begin{cases} \left[ \left( \boldsymbol{y}_i^r - \mathbb{Z}_i^r \boldsymbol{b}_i^r \right)^\top, i \in \mathscr{N}_k(\boldsymbol{U}) \right]^\top, & \text{if } r \in \mathscr{R}^{\mathsf{Num}}, \\ \left[ \left( \boldsymbol{y}_i^{\star,r} - \mathbb{Z}_i^r \boldsymbol{b}_i^r \right)^\top, i \in \mathscr{N}_k(\boldsymbol{U}) \right]^\top, & \text{if } r \in \mathscr{R}^{\mathsf{OB}}. \end{cases}$$

Using basic algebraic operations and ignoring several multiplicative constants, we can rewrite the probability density function of full-conditioned distribution of $\boldsymbol{\beta}^{(k),r}$ into:

$$p\left( \boldsymbol{\beta}^{(k),r} \middle| \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \tau_r^{(k)}; \boldsymbol{\beta}_0^r, \mathbb{D}^r, \mathscr{C}^r \right) \propto$$
$$\exp\left\{ -\frac{\tau_r^{(k)}}{2} \left( \boldsymbol{\beta}^{(k),r} - \widetilde{\boldsymbol{\beta}}^{(k),r} \right)^\top \left[ \left( \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r \right)^\top \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r + (\mathbb{D}^r)^{-1} \right] \left( \boldsymbol{\beta}^{(k),r} - \widetilde{\boldsymbol{\beta}}^{(k),r} \right) \right\},$$

where

$$\widetilde{\boldsymbol{\beta}}^{(k),r} = \left[ \left( \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r \right)^\top \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r + (\mathbb{D}^r)^{-1} \right]^{-1} \left( \left( \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r \right)^\top \widetilde{\boldsymbol{y}}_{\mathscr{N}_k(\boldsymbol{U})}^r + (\mathbb{D}^r)^{-1} \boldsymbol{\beta}_0^{(k),r} \right),$$

which compared to pdf of multivariate normal distribution yields

$$\boldsymbol{\beta}^{(k),r} \left| \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \tau_r^{(k)}; \boldsymbol{\beta}_0^r, \mathbb{D}^r, \mathscr{C}^r \right. \sim \mathsf{N}_{d_r^{\mathsf{F}}}\left( \widetilde{\boldsymbol{\beta}}^{(k),r}, \frac{1}{\tau_r^{(k)}} \left[ \left( \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r \right)^\top \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r + (\mathbb{D}^r)^{-1} \right]^{-1} \right).$$

### *A.7 Prior precisions $\boldsymbol{\tau}_{\mathsf{R}}$ for $\boldsymbol{\mu}$*

Parameter $\boldsymbol{\tau}_{\mathsf{R}}$ serves as an auxiliary parameter for specifying prior distribution of $\boldsymbol{\mu}$, see Section 4.1. The derivation of the full-conditioned distribution of this parameter is then solely based on combining $p(\boldsymbol{\mu}|\boldsymbol{\tau}_{\mathsf{R}})$ and with the prior $\Gamma(a_3,a_4)$. Therefore,

$$p\left(\boldsymbol{\tau}_{\mathsf{R}}\,|\,\boldsymbol{\mu};\boldsymbol{\mu}_0,a_3,a_4\right) \propto \prod_{k=1}^{K}\prod_{j=1}^{d^{\mathsf{R}}}\left(\tau_{\mathsf{R},j}^{(k)}\right)^{a_3+\frac{1}{2}-1}\exp\left\{-\tau_{\mathsf{R},j}^{(k)}\left[a_4+\frac{1}{2}\left(\mu_j^{(k)}-\mu_{0,j}^{(k)}\right)^2\right]\right\}$$

and

$$\tau^{(k)}_{\mathsf{R},j}\,\Big|\,\mu_j^{(k)};\mu_{0,j}^{(k)},a_3,a_4 \;\sim\; \Gamma\left(a_3+\frac{1}{2},a_4+\frac{1}{2}\left(\mu_j^{(k)}-\mu_{0,j}^{(k)}\right)^2\right)$$

independently for all $j=1,\ldots,d^{\mathsf{R}}$ and $k=1,\ldots,D$.

### *A.8 Prior expected values $\boldsymbol{\mu}$ for $\boldsymbol{b}$*

Parameter $\boldsymbol{\mu}$ consists of all possible expected values $\boldsymbol{\mu}^{(k)}$ of random effects $\boldsymbol{b}_i$ in all groups $k=1,\ldots,K$. The right-hand side of (21) is, in the case of this parameter, simplified into

$$p\left(\boldsymbol{\mu}\,|\,\mathbb{Y},\boldsymbol{\Psi}_{-\boldsymbol{\mu}};\boldsymbol{\zeta},\mathscr{C}\right)\propto\prod_{i=1}^{n}p\left(\boldsymbol{b}_i\,\Big|\,\boldsymbol{\mu}^{(U_i)},\boldsymbol{\Sigma}^{-(U_i)}\right)\cdot p\left(\boldsymbol{\mu}|\boldsymbol{\tau}_{\mathsf{R}};\boldsymbol{\mu}_0\right).$$

From the product across all subjects for given group $k=1,\ldots,K$ we extract only those factors that correspond to subjects within the $k$-th group, i.e. $\mathscr{N}_k(\boldsymbol{U})$. By performing several algebraic operations and ignoring multiplicative constants, we obtain

$$\begin{aligned}
p\left(\boldsymbol{\mu}^{(k)}\,\Big|\,\boldsymbol{U},\boldsymbol{b},\boldsymbol{\Sigma}^{-(k)},\boldsymbol{\tau}_{\mathsf{R}}^{(k)};\boldsymbol{\mu}_0^{(k)}\right) &\propto \prod_{i\in\mathscr{N}_k(\boldsymbol{U})}p\left(\boldsymbol{b}_i\,\Big|\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{-(k)}\right)\cdot p\left(\boldsymbol{\mu}^{(k)}|\boldsymbol{\tau}_{\mathsf{R}}^{(k)};\boldsymbol{\mu}_0^{(k)}\right)\\
&\propto \exp\left\{-\frac{1}{2}\sum_{i\in\mathscr{N}_k(\boldsymbol{U})}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)^{\top}\boldsymbol{\Sigma}^{-(k)}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)-\frac{1}{2}\left(\boldsymbol{\mu}^{(k)}-\boldsymbol{\mu}_0^{(k)}\right)^{\top}\mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right)\left(\boldsymbol{\mu}^{(k)}-\boldsymbol{\mu}_0^{(k)}\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}^{(k)}-\widetilde{\boldsymbol{\mu}}^{(k)}\right)^{\top}\left[n^k(\boldsymbol{U})\boldsymbol{\Sigma}^{-(k)}+\mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right)\right]\left(\boldsymbol{\mu}^{(k)}-\widetilde{\boldsymbol{\mu}}^{(k)}\right)\right\},
\end{aligned}$$

where

$$\widetilde{\boldsymbol{\mu}}^{(k)}=\left[n^k(\boldsymbol{U})\boldsymbol{\Sigma}^{-(k)}+\mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right)\right]^{-1}\left(n^k(\boldsymbol{U})\boldsymbol{\Sigma}^{-(k)}\underbrace{\frac{1}{n^k(\boldsymbol{U})}\sum_{i\in\mathscr{N}_k(\boldsymbol{U})}\boldsymbol{b}_i}_{\bar{\boldsymbol{b}}^k(\boldsymbol{U})}+\mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right)\boldsymbol{\mu}_0\right)$$

leading to the following full-conditioned distribution

$$\boldsymbol{\mu}^{(k)}\,\Big|\,\boldsymbol{U},\boldsymbol{b},\boldsymbol{\Sigma}^{-(k)},\boldsymbol{\tau}_{\mathsf{R}}^{(k)};\boldsymbol{\mu}_0^{(k)}\;\sim\;\mathsf{N}_{d^{\mathsf{R}}}\left(\widetilde{\boldsymbol{\mu}}^{(k)},\left[n^k(\boldsymbol{U})\boldsymbol{\Sigma}^{-(k)}+\mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right)\right]^{-1}\right)$$

independently for all $k=1,\ldots,K$.

### *A.9 Prior scale matrices $\mathbb{Q}^{-1}$ for $\boldsymbol{\Sigma}^{-1}$*

Parameter $\mathbb{Q}^{-1}$ is the set of auxiliary parameters that makes prior distribution of $\boldsymbol{\Sigma}^{-1}$ more flexible within Gibbs sampler. The right-hand side of (21) shrinks into

$$p\left(\mathbb{Q}^{-1}\,|\,\mathbb{Y},\boldsymbol{\Psi}_{-\mathbb{Q}^{-1}};\boldsymbol{\zeta},\mathscr{C}\right)\propto p\left(\boldsymbol{\Sigma}^{-1}|\,\mathbb{Q}^{-1};\nu_0\right)\cdot p\left(\mathbb{Q}^{-1}|\,\mathbb{D}^{\mathbb{Q}},\nu_1\right),$$

where the two pdfs on the right hand side correspond to Wishart distribution. Combining them we get

$$p\left(\mathbb{Q}^{-(k)}\,\middle|\,\boldsymbol{\Sigma}^{-(k)};\nu_0,\nu_1,\mathbb{D}^{\mathbb{Q}}\right) \propto \left|\mathbb{Q}^{-(k)}\right|^{\frac{\nu_0+\nu_1-d^{\mathsf{R}}-1}{2}} \exp\left\{-\operatorname{\mathsf{Tr}}\left[\left(\boldsymbol{\Sigma}^{-(k)}+\left(\mathbb{D}^{\mathbb{Q}}\right)^{-1}\right)\mathbb{Q}^{-(k)}\right]\right\},$$

which resembles pdf of Wishart distribution. Therefore,

$$\mathbb{Q}^{-(k)}\,\middle|\,\boldsymbol{\Sigma}^{-(k)};\nu_0,\nu_1,\mathbb{D}^{\mathbb{Q}} \;\sim\; \mathsf{W}_{d^{\mathsf{R}}}\left(\left[\boldsymbol{\Sigma}^{-(k)}+\left(\mathbb{D}^{\mathbb{Q}}\right)^{-1}\right]^{-1},\nu_0+\nu_1\right)$$

independently for all $k=1,\dots,K$.

### *A.10 Prior inverse covariance matrices* $\boldsymbol{\Sigma}^{-1}$ *for random effects* $\boldsymbol{b}$

Parameter $\boldsymbol{\Sigma}^{-1}$ is the set of inverse covariance matrices for random effects $\boldsymbol{b}_i$ that contributes to the right-hand side of (21) only in the pdf for random effects and in the prior distribution of $\boldsymbol{\Sigma}^{-1}$:

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

Again, we need to separate subjects into the groups $\mathscr{N}_k(\boldsymbol{U}), k=1,\dots,K$ according to their current allocation indicators $\boldsymbol{U}$. Similar to before, the equation above decomposes into $K$ independent parts - one for each group $k=1,\dots,K$. Considering the $k$th group, the right-hand side of the equation above reduces into

$$\begin{aligned}
p\left(\boldsymbol{\Sigma}^{-(k)}\,\middle|\,\boldsymbol{U},\boldsymbol{b},\boldsymbol{\mu}^{(k)},\mathbb{Q}^{-(k)};\nu_0\right) &\propto \left|\boldsymbol{\Sigma}^{-(k)}\right|^{\frac{n^k(\boldsymbol{U})+\nu_0-d^{\mathsf{R}}-1}{2}}\cdot\\
&\quad\cdot\exp\left\{-\frac{1}{2}\sum_{i\in\mathscr{N}_k(\boldsymbol{U})}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)^{\top}\boldsymbol{\Sigma}^{-(k)}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)-\operatorname{\mathsf{Tr}}\left[\mathbb{Q}^{-(k)}\boldsymbol{\Sigma}^{-(k)}\right]\right\}\\
&\propto \left|\boldsymbol{\Sigma}^{-(k)}\right|^{\frac{n^k(\boldsymbol{U})+\nu_0-d^{\mathsf{R}}-1}{2}}\exp\left\{-\operatorname{\mathsf{Tr}}\left[\left(\mathbb{Q}^{-(k)}+\frac{1}{2}\sum_{i\in\mathscr{N}_k(\boldsymbol{U})}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)^{\top}\right)\boldsymbol{\Sigma}^{-(k)}\right]\right\},
\end{aligned}$$

which again resembles the pdf of Wishart distribution. Therefore, independently for all $k=1,\dots,K$

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

where

$$\widetilde{\mathbb{Q}}^{(k)}=\left(\widetilde{\mathbb{Q}}^{-(k)}\right)^{-1} \quad\text{and}\quad \widetilde{\mathbb{Q}}^{-(k)}=\mathbb{Q}^{-(k)}+\frac{1}{2}\sum_{i\in\mathscr{N}_k(\boldsymbol{U})}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)^{\top}.$$

### *A.11 Random effects* $\boldsymbol{b}$

The key role of our model is played by the random effects $\boldsymbol{b}_i$, $i=1,\dots,n$ that create linear predictors $\boldsymbol{\eta}_i^{(k),r}$, $k=1,\dots,K$ and $r\in\mathscr{R}$. The probability density function of corresponding full-conditioned distribution is based on only two parts of the right-hand side of (21):

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

Clearly, random effects $\boldsymbol{b}_i$ will be distributed independently even in the full-conditioned distribution. Let us select subject $i$ (say from group $U_i=k$), in which case its corresponding pdf is of the shape

$$\begin{aligned}
p\left(\boldsymbol{b}_i\,\middle|\,\mathbb{Y}_i^{\mathsf{N}},\mathbb{Y}_i^{\star,\mathsf{OB}},U_i,\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1};\mathscr{C}_i\right) &\propto \prod_{r\in\mathscr{R}^{\mathsf{Num}}}\exp\left\{-\frac{\tau_r^{(k)}}{2}\left(\widetilde{\boldsymbol{y}}_i^{r}-\mathbb{Z}_i^{r}\boldsymbol{b}_i^{r}\right)^{\top}\left(\widetilde{\boldsymbol{y}}_i^{r}-\mathbb{Z}_i^{r}\boldsymbol{b}_i^{r}\right)\right\}\cdot\\
&\quad\cdot\prod_{r\in\mathscr{R}^{\mathsf{OB}}}\exp\left\{-\frac{1}{2}\left(\widetilde{\boldsymbol{y}}_i^{\star,r}-\mathbb{Z}_i^{r}\boldsymbol{b}_i^{r}\right)^{\top}\left(\widetilde{\boldsymbol{y}}_i^{\star,r}-\mathbb{Z}_i^{r}\boldsymbol{b}_i^{r}\right)\right\}\cdot\exp\left\{-\frac{1}{2}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)^{\top}\boldsymbol{\Sigma}^{-(k)}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)\right\},
\end{aligned}$$

where $\widetilde{\boldsymbol{y}}_i^r = \boldsymbol{y}_i^r - \mathbb{X}_i^r \boldsymbol{\beta}^r$ and $\widetilde{\boldsymbol{y}}_i^{\star,r} = \boldsymbol{y}_i^{\star,r} - \mathbb{X}_i^r \boldsymbol{\beta}^r$. Constructing

$$\widetilde{\boldsymbol{y}}_i = \left(\begin{array}{c} \vdots \\ \sqrt{\tau_r^{(k)}}\widetilde{\boldsymbol{y}}_i^r \\ \vdots \\ \widetilde{\boldsymbol{y}}_i^{\star,r} \\ \vdots \end{array}\right) \begin{array}{l} r \in \mathscr{R}^{\mathsf{Num}}, \\ \\ r \in \mathscr{R}^{\mathsf{OB}}, \end{array} , \qquad \widetilde{\mathbb{Z}}_i = \begin{pmatrix} \ddots & & & \\ & \sqrt{\tau_r^{(k)}}\mathbb{Z}_i^r & & \\ & & \ddots & \\ & & & \mathbb{Z}_i^r \\ & & & & \ddots \end{pmatrix}$$

we can simplify the above to

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

which after several algebraic operations and ignoring multiplicative constants becomes

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

where

$$\widetilde{\boldsymbol{b}}_i = \left[\widetilde{\mathbb{Z}}_i^\top \widetilde{\mathbb{Z}}_i + \boldsymbol{\Sigma}^{-(k)}\right]^{-1} \left(\widetilde{\mathbb{Z}}_i^\top \widetilde{\boldsymbol{y}}_i + \boldsymbol{\Sigma}^{-(k)}\boldsymbol{\mu}^{(k)}\right).$$

Therefore, the full-conditioned distribution of $\boldsymbol{b}_i$ for a subject belonging to group $k = 1, \ldots, K$ is

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

## References

Aitkin M, Liu CC, Chadwick T (2009) Bayesian model comparison and model averaging for small-area estimation. The Annals of Applied Statistics 3(1):199 – 221, DOI 10.1214/08-AOAS205

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88(422):669–679, DOI 10.2307/2290350

Banfield D J, Raftery E A (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49(3):803–821

Brooks S, Gelman A, Jones G, Meng X (2011) Handbook for Markov chain Monte Carlo, 2nd edn. Taylor & Francis

Bruckers L, Molenberghs G, Drinkenburg P, Geys H (2016) A clustering algorithm for multivariate longitudinal data. Journal of Biopharmaceutical Statistics 26(4):725–741

Celeux G, Martin O, Lavergne C (2005) Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. Statistical Modelling 5(3):243–267, DOI 10.1191/1471082X05st096oa

De la Cruz-Mesía R, Quintana FA, Marshall G (2008) Model-based clustering for longitudinal data. Computational Statistics and Data Analysis 52(3):1441–1457, DOI 10.1016/j.csda.2007.04.005

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39(1):1–38

Fieuws S, Verbeke G (2004) Joint modelling of multivariate longitudinal profiles: Pitfalls of the random-effects approach. Statistics in medicine 23:3093–3104, DOI 10.1002/sim.1885

Fieuws S, Verbeke G (2006) Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. Biometrics 62(2):424–431

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97(458):611–631, DOI 10.1198/016214502760047131

Frühwirth-Schnatter S (2006) Finite Mixture and Markov Switching Models. Springer

Frühwirth-Schnatter S (2011) Panel data analysis: A survey on model-based clustering of time series. Advances in Data Analysis and Classification 5(4):251–280, DOI 10.1007/s11634-011-0100-0

Frühwirth-Schnatter S, Malsiner-Walli G (2019) From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. Advances in Data Analysis and Classification 13(1):33–64, DOI 10.1007/s11634-018-0329-y

Frühwirth-Schnatter S, Pamminger C, Weber A, Winter-Ebmer R (2012) Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. Journal of Applied Econometrics 27:1116–1137, DOI 10.1002/jae.1249

Frühwirth-Schnatter S, Pittner S, Weber A, Winter-Ebmer R (2018) Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering. The Annals of Applied Statistics 12:1796–1830, DOI 10.1214/17-AOAS1132

Genz A (1992) Numerical computation of multivariate normal probabilities. Journal of Computational and Graphical Statistics 1(2):141–149

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2019) mvtnorm: Multivariate Normal and t Distributions. URL https://CRAN.R-project.org/package=mvtnorm, R package version 1.0-11

Grün B, Leisch F (2008) FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. Journal of Statistical Software 28(4):1–35, DOI 10.18637/jss.v028.i04

Grün B (2019) Model-based clustering. In: Frühwirth-Schnatter S, Celeux G, Robert CP (eds) Handbook of Mixture Analysis, CRC Press, chap 8, pp 157–192

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second edn. Springer Science+Business Media, New York

James GM, Sugar CA (2003) Clustering for sparsely sampled functional data. Journal of the American Statistical Association 98(462):397–408, DOI 10.1198/016214503000189

Komárek A, Komárková L (2013) Clustering for multivariate continuous and discrete longitudinal data. The Annals of Applied Statistics 7(1):177–200, DOI 10.1214/12-AOAS580

Komárek A, Komárková L (2014) Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. Journal of Statistical Software 59(12):1–38, DOI 10.18637/jss.v059.i12

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38(4):963–974

Liu X, Yang MCK (2009) Simultaneous curve registration and clustering for functional data. Computational Statistics and Data Analysis 53(4):1361–1376, DOI 10.1016/j.csda.2008.11.019

Ma P, Castillo-Davis CI, Zhong W, Liu JS (2006) A data-driven clustering method for time course gene expression data. Nucleic Acids Research 34(4):1261–1269, DOI 10.1093/nar/gkl013

McNicholas PD, Murphy TB (2010) Model-based clustering of longitudinal data. The Canadian Journal of Statistics 38(1):153–168, DOI 10.1002/cjs.10047

Molenberghs G, Verbeke G (2005) Models for Discrete Longitudinal Data. Springer, New York

Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9(2):249–265

Proust-Lima C, Philipps V, Diakite A, Liquet B (2017) Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. Journal of Statistical Software 78(2):1–56, DOI doi:10.18637/jss.v078.i02

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org

Raftery AE, Dean N (2006) Variable selection for model-based clustering. Journal of the American Statistical Association 101(473):168–178

Stephens M (2000) Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62(4):795–809

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82(398):528–550, DOI 10.2307/2289457

Verbeke G, Lesaffre E (1996) A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association 91(433):217–221, DOI 10.1080/01621459.1996.10476679

Villarroel L, Marshall G, Barón AE (2009) Cluster analysis using multivariate mixed effects models. Statistics in Medicine 28(20):2552–2565, DOI 10.1002/sim.3632