Advances in Data Analysis and Classification manuscript No.
(will 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: June 1, 2021 / Accepted: Month DD, YYYY

Abstract Although, many nowadays studies gather data of diverse nature (numeric quantities, binary indicators or ordered categories) on the same units repeatedly over time, there exist only limited number of approaches in the literature to analyse so called *mixed-type* longitudinal data. We present a statistical model capable of joint modelling of several mixed-type outcomes, which also accounts for possible dependencies among 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 (MBC) methodology to create a mixture of such models to model heterogeneity in temporal evolution of considered outcomes. The estimation of such hierarchical model is approached by Bayesian principles with the use of Markov Chain Monte Carlo (MCMC) methods. After a successful simulation study with the aim to examine the ability to consistently estimate the true parameter values and thus discovering the different patterns, the EU-SILC dataset consisting of Czech households each followed for four years in a time span 2005 – 2016 was analysed. Households were classified into groups with 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. On top of that, 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 respective countries and variables, which are collected annually via questionnaires, refer both to households and to individuals from the household. In this paper, we concentrate on household specific data from the Czech Republic (period 2005 – 2016), each household was followed annually for a period of 4 years. In total,

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

$n = 20\,323$ households will be analyzed. The aim of the research is to find typical patterns of 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 cover unexpected expenses), or *ordinal* (e.g., level of a financial burden of housing). From a data analytic point of view, it is our aim to develop a clustering approach suitable for longitudinal data of a mixed type that allows for 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 ith unit ($i = 1,\ldots,n$), in total R outcome variables $Y^r_{i,j}$ ($r = 1,\ldots,R$, $j = 1,\ldots,n_i$) are being gathered at each of n_i measurement occasions that take place at times $t_{i,1},\ldots,t_{i,n_i}$. On top of that, each outcome variable $Y^r_{i,j}$ might be either *numeric*, *binary* or *ordinal*, see Figure 1 for an example. 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 ith unit is represented by data $\mathscr{D}_i = \left\{Y^r_{i,j}, \boldsymbol{v}^r_{i,j}, t_{i,j} : r = 1,\ldots,R,\ j = 1,\ldots,n_i\right\}$, $i = 1,\ldots,n$ 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 complexity of a data structure and also due to the fact that possibly different numbers n_i of measurement occasions appear in data for different units, classical distance-based clustering methods like 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. A classical model to analyze *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. A 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 being *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 keep dealing with *continuous* and moreover univariate ($R = 1$) longitudinal data (e.g. James and Sugar, 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

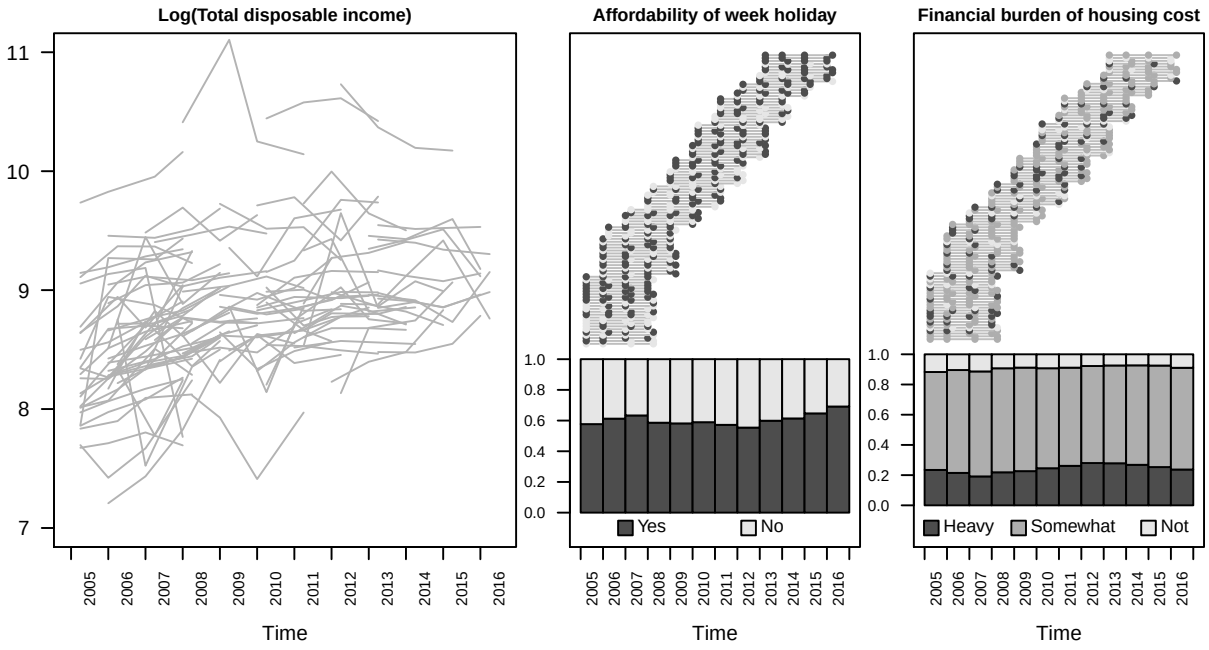


Fig. 1: EU-SILC data (Czech Republic). Observed longitudinal profiles of three (numeric, binary and ordinal) outcomes.

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 seems to be available in the literature in case 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 into 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 are lacking 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 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 a typical questionnaire data like the EU-SILC database where many outcome variables are of an ordinal nature.

One of the reasons why not much is available to perform clustering based on multivariate mixed type longitudinal data is perhaps also 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 especially if we seek for 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 also in this manuscript, we 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 covariance matrix of a high dimension. Nevertheless, MBC was not employed in any of those solutions. Recently, Bruckers et al (2016) invented a clustering algorithm that updates pseudo-log-likelihood of the pairwise approach and reclassifies individuals until no change is made. This solution, however, lacks inclusion of binary and ordinal outcomes that we aim to provide in this article.

The rest of the paper is organized as follows. In Section 2 we first outline the approach being 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 on Figure 1 and classification of study units into groups with apriori unknown structure. Yet, Section 3 only provides a theoretical clustering concept which assumes a full knowledge of unknown parameters. 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 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 to classifying each household. Finally, Section 7 summarizes the proposed model and suggests ways of improvement in reaction to our findings from application.

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 $\mathscr{R} = \{1, \ldots, R\} = \mathscr{R}^{\mathsf{Num}} \cup \mathscr{R}^{\mathsf{OB}}$, $\mathscr{R}^{\mathsf{OB}} = \mathscr{R}^{\mathsf{Ord}} \cup \mathscr{R}^{\mathsf{Bin}}$, 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 $\boldsymbol{Y}_i^r = \left(Y_{i,1}^r, \ldots, Y_{i,n_i}^r\right)^\top$ be the vector of values of outcome $r \in \mathscr{R}$ of subject $i = 1, \ldots, n$ observed at times $\boldsymbol{t}_i = (t_{i,1}, \ldots, t_{i,n_i})$ together with additional covariates $\boldsymbol{v}_{i,1}^r, \ldots, \boldsymbol{v}_{i,n_i}^r$. Further, let $\mathscr{C}_i^r = \left\{\boldsymbol{t}_i, \boldsymbol{v}_{i,1}^r, \ldots, \boldsymbol{v}_{i,n_i}^r\right\}$ denote both the measurement times and the covariate values for the outcome r of the ith subject. Finally, let

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

denote all information (outcomes and covariate values) available for the ith 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 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:

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

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, $\mathbb{X}_i^r = \left(\boldsymbol{x}_{i,1}^r, \ldots, \boldsymbol{x}_{i,n_i}^r\right)^\top$ and $\mathbb{Z}_i^r = \left(\boldsymbol{z}_{i,1}^r, \ldots, \boldsymbol{z}_{i,n_i}^r\right)^\top$ 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. created regressor falls exclusively either into fixed effects part or into random effects part of the model.

2.2 Ordinal and binary longitudinal outcomes

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

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

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 $\boldsymbol{\gamma}^r = \left(\gamma_1^r, \ldots, \gamma_{L^r-1}^r\right)$ be the vector of thresholds. For identifiability purposes, we will fix one of the thresholds, e.g., the first one γ_1^r while estimating the model. That is, in 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

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

with analogous notation to that in (2). Nevertheless, this time, the precision parameter τ_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 being 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 $\boldsymbol{b}_i = \left\{\boldsymbol{b}_i^r, r \in \mathscr{R}\right\}$.

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.

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 observed data:

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

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

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

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

3 Model based clustering framework

Classification of 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 $U_i \in \left\{1, \ldots, K\right\}$ 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 $p\left(\mathbb{Y}_i \middle| \mathscr{C}_i; \boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\gamma}\right)$ 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 model parameters. That is, the assumed conditional probability distribution function of the i-th subject outcomes given the group allocation is

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

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

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

In a majority of the MBC methodologies, the authors consider maximum-likelihood estimation (MLE) of the unknown parameters. 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. In our situation, it would maximize the likelihood function:

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

This is traditionally solved by using the EM algorithm (Dempster et al, 1977) to face a problem of latent allocations leading to the mixture type likelihood. Nevertheless, two other levels of latent variables are present in our model leading to two additional levels of integration when evaluation the likelihood, see expression (6). They are the random effects $\boldsymbol{b}_i$ and the latent numeric variables $\mathbb{Y}_i^{\star,\mathsf{OB}}$ associated with ordinal and binary outcomes $\mathbb{Y}_i^{\mathsf{OB}}$ as well. This makes the likelihood hardly tractable and we switch to the Bayesian framework and the related Markov chain Monte Carlo (MCMC) methodology which allows to fully exploit a hierarchical structure of our model. The clustering itself will then be based on the posterior distribution of the individual group probabilities (9).

4 Bayesian inference

For Bayesian inference, we exploit ideas of Bayesian Data Augmentation (BDA, Tanner and Wong, 1987) while considering all latent quantities, i.e., component allocations $\boldsymbol{U} := \{U_i, i = 1, \ldots, n\}$, LMM random effect vectors $\boldsymbol{b} := \{\boldsymbol{b}_i, i = 1, \ldots, n\}$ and latent variables $\mathbb{Y}^{\star,\mathsf{OB}} := \{\mathbb{Y}_i^{\star,\mathsf{OB}}, i = 1, \ldots, n\}$ 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

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

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 a collection of corresponding outcomes of the same type across all individuals, i.e. $\mathbb{Y}^{\mathsf{N}} = \{\mathbb{Y}_i^{\mathsf{N}}, i = 1, \ldots, n\}$ and $\mathbb{Y}^{\mathsf{OB}} = \{\mathbb{Y}_i^{\mathsf{OB}}, i = 1, \ldots, n\}$. Later, we also use $\mathbb{Y}$ for all outcomes available, that is $\mathbb{Y} = \mathbb{Y}^{\mathsf{N}} \cup \mathbb{Y}^{\mathsf{OB}}$.

4.1 Prior distribution

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

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

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 tackle the identifiability issue. Corresponding parametric space Ω^r is limited to set of all vectors of ordered values with fixed first threshold γ_1^r. An improper uniform distribution on Ω^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 $\boldsymbol{\beta}^{r,(k)} = \left(\beta_1^{r,(k)}, \ldots, \beta_{d_r^{\mathsf{F}}}^{r,(k)}\right)$, $r \in \mathscr{R}$, $k = 1, \ldots, K$, are assumed to be apriori independent and following a conjugate normal distributions, i.e.,

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

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.,

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

where each $p\left(\tau_r^{(k)}\right)$ corresponds to the gamma distribution $\Gamma(a_1, a_2)$.

Also for the random effect means, a set of independent, for simplicity only semi-conjugate normal priors is assumed, i.e.,

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

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

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 ν_1. In our applications we use $\nu_0 = \nu_1 = d^{\mathsf{R}} + 1$ and $\mathbb{D}^{\mathbb{Q}} = 100 \cdot \mathbb{I}_{d^{\mathsf{R}}}$.

4.2 MCMC sampling scheme

Posterior distribution $p(\boldsymbol{\theta} | \mathbb{Y}, \mathscr{C})$ of model parameters $\boldsymbol{\theta}$ and its characteristics will be estimated using MCMC methodology Brooks et al (2011). We adopted well known Gibbs sampling scheme which samples a new value of each of the parameters from its full-conditioned distribution while always utilizing the last known values of other parameters. Due to our (semi)-conjugate choice of prior distributions the full-conditioned distributions are from well-known families, thus straightforward to be sampled from. Derivations of all full-conditioned distributions are postponed to Appendix section A. In this section we present only the set of required full-conditioned distributions for Gibbs sampling algorithm using the notation declared in Appendix:

1. $\boldsymbol{w} | \boldsymbol{U} \overset{(37)}{\sim} \mathsf{Dir}_K\left(\boldsymbol{n}(\boldsymbol{U}) + \boldsymbol{\alpha}\right)$,
2. $\mathsf{P}\left(U_i = k \,\middle|\, \mathbb{Y}_i^{\mathsf{N}}, \mathscr{C}_i; \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{w}\right) \overset{(38)}{=} \cdots$,
3. $Y_{i,j}^{\star,r} \,\middle|\, Y_{i,j}^r = l, \boldsymbol{\gamma} \overset{(39)}{\sim} \mathsf{TN}\left(\eta_{i,j}^{(U_i),r}, 1, \gamma_l^r, \gamma_{l+1}^r\right) \qquad \text{for } r \in \mathscr{R}^{\mathsf{OB}}, l = 0, \ldots, L^r - 1,$
4. $\gamma_l^r \,|\boldsymbol{Y}^r; \boldsymbol{Y}^{\star,r} \overset{(40)}{\sim} \mathsf{Unif}\left[\max_{y \in \mathscr{Y}_{l-1}^r} y, \min_{y \in \mathscr{Y}_l^r} y\right] \qquad \text{for } r \in \mathscr{R}^{\mathsf{Ord}}, l = 2, \ldots, L^r - 1,$
5. $\tau_r^{(k)} \,\middle|\, \boldsymbol{Y}^r, \mathscr{C}^r; \boldsymbol{U}, \boldsymbol{b}^r; \boldsymbol{\beta}^{(k),r} \overset{(42)}{\sim} \Gamma\left(\widetilde{a}_1^{(k),r}, \widetilde{a}_2^{(k),r}\right) \qquad \text{for } r \in \mathscr{R}^{\mathsf{Num}},$
6. $\boldsymbol{\beta}^{(k),r} \,\middle|\, \boldsymbol{Y}^r, \mathscr{C}^r; \boldsymbol{U}, \boldsymbol{b}^r; \tau_r^{(k)} \overset{(44)}{\sim} \mathsf{N}_{d_r^{\mathsf{F}}}\left(\widetilde{\boldsymbol{\beta}}^{(k),r}, \left(\tau_r^{(k)}\right)^{-1} \left[\left(\mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r\right)^{\top} \mathbb{X}_{\mathscr{N}_k(\boldsymbol{U})}^r + (\mathbb{D}^r)^{-1}\right]^{-1}\right) \qquad \text{for } r \in \mathscr{R},$
7. $\tau_{\mathsf{R},j}^{(k)} \,\middle|\, \mu_j^{(k)} \overset{(46)}{\sim} \Gamma\left(a_3 + \tfrac{1}{2}, a_4 + \frac{1}{2}\left(\mu_j^{(k)} - \mu_{0,j}^{(k)}\right)^2\right),$
8. $\boldsymbol{\mu}^{(k)} \,\middle|\, \boldsymbol{U}, \boldsymbol{b}; \boldsymbol{\Sigma}^{(k)}; \boldsymbol{\tau}_{\mathsf{R}}^{(k)} \overset{(48)}{\sim} \mathsf{N}_{d^{\mathsf{R}}}\left(\widetilde{\boldsymbol{\mu}}^{(k)}, \left[n^k(\boldsymbol{U})\boldsymbol{\Sigma}^{-(k)} + \mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right)\right]^{-1}\right),$

9. $\mathbb{Q}^{-(k)} \left| \boldsymbol{\Sigma}^{(k)} \overset{(50)}{\sim} \mathsf{W}_{d^{\mathsf{R}}} \left(\left[\boldsymbol{\Sigma}^{-(k)} + \left(\mathbb{D}^{\mathbb{Q}} \right)^{-1} \right]^{-1}, \nu_0 + \nu_1 \right) \right.$,
10. $\boldsymbol{\Sigma}^{-(k)} \left| \boldsymbol{U}, \boldsymbol{b}; \boldsymbol{\mu}^{(k)}, \mathbb{Q}^{(k)} \overset{(52)}{\sim} \mathsf{W}_{d^{\mathsf{R}}} \left(\widetilde{\mathbb{Q}}^{(k)}, n^k(\boldsymbol{U}) + \nu_0 \right) \right.$,
11. $\boldsymbol{b}_i \left| \mathbb{Y}_i^{\mathsf{N}}, \mathscr{C}_i; \mathbb{Y}_i^{\star,\mathsf{OB}}, U_i; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \overset{(54)}{\sim} \mathsf{N}_{d^{\mathsf{R}}} \left(\widetilde{\boldsymbol{b}}_i, \left[\widetilde{\mathbb{Z}}_i^{\top} \widetilde{\mathbb{Z}}_i + \boldsymbol{\Sigma}^{-(k)} \right]^{-1} \right) \right.$,

A sample from converged chain constructed in this way then can be used for estimation of marginal posterior distributions of individual parameters.

4.3 Classification probabilities

Posterior distribution is invariant towards permutation of cluster labels. Hence, before calculation of classification probabilities we consider problem of *label switching* by method of Stephens (2000). This post-sampling procedure that considers all $K!$ permutations of labels for each iteration ensures that the latent clusters $1, \ldots, K$ have fixed meaning during the whole sampling procedure. However, in our applications no need of permutation has occurred.

Primarily, we perform classification using posterior means $\widehat{U}_{i,k} = \int_{\boldsymbol{\theta}} u_{i,k}(\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} | \mathbb{Y}, \mathscr{C}) \, \mathrm{d}\boldsymbol{\theta}$ of allocation probabilities $u_{i,k}(\boldsymbol{\theta})$ defined in (9). With MCMC based inference we need to calculate $u_{i,k}(\boldsymbol{\theta})$ at each iteration of sampled $\boldsymbol{\theta}$, for which probability density functions of i-th outcomes given cluster need to be evaluated (8). Unfortunately, this involves non-trivial integration of auxiliary latent variables - random effects $\boldsymbol{b}_i$ and latent numeric outcomes $\mathbb{Y}_i^{\star,\mathsf{OB}}$. Therefore, we devote the rest of this section to description of a method chosen for calculating desired integrals

$$
\begin{aligned}
& p\left(\mathbb{Y}_i \left| U_i = k, \mathscr{C}_i; \, \boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\gamma} \right. \right) = \\
& \quad = \int p\left(\mathbb{Y}_i^{\mathsf{OB}} \left| \mathbb{Y}_i^{\star,\mathsf{OB}}; \boldsymbol{\gamma} \right. \right) \cdot \underbrace{\left[\int p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \left| \mathscr{C}_i, \boldsymbol{b}_i; \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)} \right. \right) \cdot p\left(\boldsymbol{b}_i \left| \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)} \right. \right) \mathrm{d}\boldsymbol{b}_i \right]}_{p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \left| \mathscr{C}_i; \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)} \right. \right)} \mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}. \qquad (11)
\end{aligned}
$$

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

Let us first perform the integration of random effects $\boldsymbol{b}_i$ from (11) to obtain distribution of numeric variables unconditioned by random effects. We will avoid integration by realization that under normality assumption of both numeric outcomes and random effects the unconditioned distribution of 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

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

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 variance matrix $\mathbb{T}$ is diagonal due to independence assumption and contains corresponding variance, that is τ_r^{-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 notoriously known formulas we obtain unconditioned mean and variance matrix:

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

which results in

$$
\boldsymbol{Y}_i \left| \mathscr{C}_i; \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)} \right. \sim \mathsf{N}_d \left(\mathbb{X}_i \boldsymbol{\beta} + \mathbb{Z}_i \boldsymbol{\mu}, \mathbb{V} \right). \qquad (12)
$$

This distribution has general covariance structure which reflects the general structure of $\boldsymbol{\Sigma}$ and hence captures dependencies among outcomes.

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

It remains to perform the following integration:

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

which is in fact integration of multivariate normal density within the bounds given by thresholds $\boldsymbol{\gamma}$ and observed ordinal and binary outcomes. First, we separate marginal distribution of numeric outcomes $\mathbb{Y}_i^{\mathsf{N}}$ since it can avoid 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:

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

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

It remains to integrate the product of two functions, first one of which just declares lower and upper integration bounds and the second one is the probability density function of multivariate normal distribution with mean $\boldsymbol{\eta}_{\mathsf{OB}}^{(k)}$ and variance 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 γ 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 $\square\left(\mathbb{Y}_i^{\mathsf{OB}}\right)=(\boldsymbol{e}_i,\boldsymbol{f}_i]\subset\mathbb{R}^{d^{\mathsf{OB}}}$ then the remaining integral can be written in the form

$$I_k(\square)=\int\limits_{\square\left(\mathbb{Y}_i^{\mathsf{OB}}\right)}p\left(\mathbb{Y}_i^{\star,\mathsf{OB}}\middle|\,\mathbb{Y}_i^{\mathsf{N}},\mathscr{C}_i;\boldsymbol{\beta}^{(k)},\boldsymbol{\tau}^{(k)},\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}\right)\mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}=\int\limits_{\boldsymbol{e}_i}^{\boldsymbol{f}_i}\varphi\left(\boldsymbol{y};\boldsymbol{\eta}_{\mathsf{OB}}^{(k)},\mathbb{V}_{\mathsf{OB}}^{(k)}\right)\mathrm{d}\boldsymbol{y}.\tag{13}$$

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

$$u_{i,k}(\boldsymbol{\theta})=\frac{w_k\cdot p\left(\mathbb{Y}_i^{\mathsf{N}}\middle|\,\mathscr{C}_i;\boldsymbol{\beta}^{(k)},\boldsymbol{\tau}^{(k)},\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}\right)\cdot I_k(\square)}{\sum\limits_{k'=1}^{K}w_{k'}\cdot p\left(\mathbb{Y}_i^{\mathsf{N}}\middle|\,\mathscr{C}_i;\boldsymbol{\beta}^{(k')},\boldsymbol{\tau}^{(k')},\boldsymbol{\mu}^{(k')},\boldsymbol{\Sigma}^{(k')}\right)\cdot I_{k'}(\square)}.\tag{14}$$

In order to compute integrals (13) needed in (14) we adopted an effective algorithm presented by Genz (1992) which is also based on MCMC sampling. Since the approximation of such an integral is needed K-times for each generated state of our Gibbs sampling, the procedure is considerably time-consuming. The implemented function `pmvnorm` in  package `mvtnorm` (Genz et al, 2019) is being used in our applications.

4.4 Classification rules

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

In order to prevent misclassification, we allow subjects to remain unclassified when the decision is not clear. One way to accomplish that would be to classify 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 chosen threshold. However, the choice of the value of this threshold for different values of K would be another problem to be dealt with. Therefore, in our applications we make use of 95% Highest Posterior Density (HPD) interval estimates. We classify a subject i into the class k with the highest classification probability $\widehat{U}_{i,k}$ if its *lower* 95% HPD bound is still higher than any other *upper* 95% HPD bound of the remaining probabilities. Otherwise, the subject i remained unclassified. That should fill clusters with theirs 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 associated with the most.

4.5 Number of groups

Throughout the paper we treated the number of latent classes K fixed as to be selected by statistician 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 try several values of K and choose the one that optimizes one of the known criteria. We follow the steps of Aitkin et al (2009) and use a deviance criterion that is solely based on the goodness of fit measured by the log-likelihood function.

Deviance is a parametric function of $\boldsymbol{\theta}$ generally defined as

$$D(\boldsymbol{\theta};\mathbb{Y},\mathscr{C}) = -2\log p(\mathbb{Y}|\mathscr{C};\boldsymbol{\theta}) = -2\sum_{i=1}^{n}\log p(\mathbb{Y}_i|\mathscr{C}_i;\boldsymbol{\theta}). \tag{15}$$

Contribution of individual i to the deviance in our model can be expressed as

$$-2\log p(\mathbb{Y}_i|\mathscr{C}_i;\boldsymbol{\theta}) = -2\log\left[\sum_{k=1}^{K}\int\int p\left(\mathbb{Y}_i,\mathbb{Y}_i^{\star,\mathsf{OB}},\boldsymbol{b}_i,U_i=k\middle|\mathscr{C}_i;\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{b}_i\,\mathrm{d}\mathbb{Y}_i^{\star,\mathsf{OB}}\right] \tag{16}$$

which includes the integration (11) for calculating classification probabilities which has been described in section 4.7. Using the same notation we can write

$$D(\boldsymbol{\theta};\mathbb{Y},\mathscr{C}) = -2\sum_{i=1}^{n}\log\left[\sum_{k=1}^{K}w_k\cdot p\left(\mathbb{Y}_i^{\mathsf{N}}\middle|\mathscr{C}_i;\boldsymbol{\beta}^{(k)},\boldsymbol{\tau}^{(k)},\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}\right)\cdot I_k\left(\square\left(\mathbb{Y}_i^{\mathsf{OB}}\right)\right)\right] \tag{17}$$

where the denominator of (14) is inserted into the logarithm. Therefore, calculation of deviance for one set of parameters $\boldsymbol{\theta}$ requires calculation of classification probabilities for every individual. Only then the deviance parametric function can be fully evaluated. Hence, exploration of posterior distribution of deviance is heavily time-consuming. This is why we applied thinning in our applications to speed up calculations of all of the probabilities.

4.6 MCMC sampling scheme

Joint density of observed outcomes $\mathbb{Y}_i$ of ith subject and its latent variables U_i, $\mathbb{Y}_i^{\star,\mathsf{OB}}$ and $\boldsymbol{b}_i$ decomposes into product of several conditional probability density functions:

$$p\left(\mathbb{Y}_i, U_i, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{b}_i \middle| \mathscr{C}_i; \boldsymbol{\theta}\right) = \\ p\left(\mathbb{Y}_i^{\mathsf{OB}} \middle| \mathbb{Y}_i^{\star,\mathsf{OB}}; \boldsymbol{\gamma}\right) \cdot p\left(\mathbb{Y}_i^{\mathsf{N}}, \mathbb{Y}_i^{\star,\mathsf{OB}} \middle| \mathscr{C}_i, \boldsymbol{b}_i; \boldsymbol{\beta}^{(U_i)}, \boldsymbol{\tau}^{(U_i)}\right) \cdot p\left(\boldsymbol{b}_i \middle| \boldsymbol{\mu}^{(U_i)}, \boldsymbol{\Sigma}^{-(U_i)}\right) \cdot p\left(U_i | \boldsymbol{w}\right), \quad (18)$$

which is easier to work with than $p(\mathbb{Y}_i|\mathscr{C}_i;\boldsymbol{\theta})$ obtainable by combining (**??**) and (8). We adopt Bayesian framework in which we consider parameters $\boldsymbol{\theta}$ to be of random nature by assigning a prior distribution $p(\boldsymbol{\theta})$ of parameter $\boldsymbol{\theta}$. Then complete probability density function of all measured or latent data (across all subjects $i = 1,\ldots,n$) and parameters is of the form:

$$p\left(\mathbb{Y}, \boldsymbol{U}, \mathbb{Y}^{\star,\mathsf{OB}}, \boldsymbol{b}, \boldsymbol{\theta} \middle| \mathscr{C}\right) = \left[\prod_{i=1}^{n} p\left(\mathbb{Y}_i, U_i, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{b}_i \middle| \mathscr{C}_i; \boldsymbol{\theta}\right)\right] \cdot p(\boldsymbol{\theta}),$$

to which is posterior probability density function $p\left(\boldsymbol{U}, \mathbb{Y}^{\star,\mathsf{OB}}, \boldsymbol{b}, \boldsymbol{\theta} \middle| \mathbb{Y}, \mathscr{C}\right)$ proportional by the Bayes theorem. Ideally, we would be interested in marginal posterior distribution of parameter $\boldsymbol{\theta}$ given by

$$p\left(\boldsymbol{\theta} \middle| \mathbb{Y}, \mathscr{C}\right) = \int\int\int p\left(\boldsymbol{U}, \mathbb{Y}^{\star,\mathsf{OB}}, \boldsymbol{b}, \boldsymbol{\theta} \middle| \mathbb{Y}, \mathscr{C}\right) \mathrm{d}\boldsymbol{b}\, \mathrm{d}\mathbb{Y}^{\star,\mathsf{OB}}\, \mathrm{d}\boldsymbol{U}.$$

To avoid computation of needed multiplicative constant and integration with respect to all latent variables we exploit MCMC methods (Robert, 2001) based on sampling Markov chain of states corresponding to parameter $\boldsymbol{\theta}$ and latent variables. As the chain reaches its stationary distribution (equal to the posterior one) generated states are considered to be sampled representatives of that distribution. Marginal posterior distribution of $\boldsymbol{\theta}$ can then be described by M generated values $\boldsymbol{\theta}_{[B+1]}, \ldots, \boldsymbol{\theta}_{[B+M]}$ after the burn-in period B.

For our model we decided to choose basic Gibbs sampling method (Geman and Geman, 1984) for constructing Markov chain of required properties. Such an algorithm requires effortless sampling from so-called *full-conditioned* distributions. We can obtain a well-known distribution when prior distributions of all components of $\boldsymbol{\theta}$ are suitably set to achieve conjugacy.

4.6.1 Prior distributions

decomposed into several independently distributed blocks: $\boldsymbol{w}$, $\boldsymbol{\gamma}$, $(\boldsymbol{\beta}, \boldsymbol{\tau})$, $(\boldsymbol{\mu}, \boldsymbol{\tau}_{\mathsf{R}})$, $\left(\boldsymbol{\Sigma}^{-1}, \mathbb{Q}^{-1}\right)$, which results in the following form of the probability density function:

$$p\left(\boldsymbol{\theta}, \mathbb{Q}^{-1}, \boldsymbol{\tau}_{\mathsf{R}} \middle| \boldsymbol{\zeta}\right) = p(\boldsymbol{w}|\boldsymbol{\zeta}) \cdot p(\boldsymbol{\gamma}|\boldsymbol{\zeta}) \cdot p(\boldsymbol{\beta}|\boldsymbol{\tau};\boldsymbol{\zeta}) \cdot p(\boldsymbol{\tau}|\boldsymbol{\zeta}) \cdot p(\boldsymbol{\mu}|\boldsymbol{\tau}_{\mathsf{R}};\boldsymbol{\zeta}) \cdot p\left(\boldsymbol{\tau}_{\mathsf{R}} \middle| \boldsymbol{\zeta}\right) \cdot p(\boldsymbol{\Sigma}^{-1}|\mathbb{Q}^{-1};\boldsymbol{\zeta}) \cdot p(\mathbb{Q}^{-1}|\boldsymbol{\zeta}),$$

where $\boldsymbol{\tau}^{\mathsf{R}}$, $\mathbb{Q}^{-1}$ are another random parameters (or set of parameters in case of group-specificity) making prior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$ more flexible and $\boldsymbol{\zeta}$ stands for set of other hyper-parameters that need to be fixed.

Since $\boldsymbol{w}$ is a vector of probabilities that add up to 1, ideal choice for prior distribution is Dirichlet distribution $\mathsf{Dir}_K(\boldsymbol{\alpha})$ with pdf equal to

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} w_k^{\alpha_k - 1}, \quad (19)$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is a set of positive parameters (set to 1 by default).

For each ordinal variable $r^{\mathsf{O}} \in \mathscr{R}^{\mathsf{O}}$ we suppose that thresholding constants $\boldsymbol{\gamma}^{r^{\mathsf{O}}}$ follow (independently) an improper uniform distribution on $\mathbb{R}^{L^{r^{\mathsf{O}}}-2}_{\gamma_1^{r^{\mathsf{O}}} <}$, i.e.

$$p\left(\boldsymbol{\gamma} \middle| \gamma_1^{r^{\mathsf{O}}}, r^{\mathsf{O}} \in \mathscr{R}^{\mathsf{Ord}}\right) \propto \prod_{r^{\mathsf{O}} \in \mathscr{R}^{\mathsf{Ord}}} \mathbb{1}_{\mathbb{R}^{L^{r^{\mathsf{O}}}-2}_{\gamma_1^{r^{\mathsf{O}}} <}}\left(\boldsymbol{\gamma}^{r^{\mathsf{O}}}\right). \quad (20)$$

For each outcome $r \in \mathscr{R}$ we suppose that fixed effects

$$\boldsymbol{\beta}^{(k),r} \left| \tau_r^{(k)}; \boldsymbol{\beta}_0^r, \mathbb{D}^r \right. \sim \mathsf{N}_{d_r^{\mathsf{F}}} \left(\boldsymbol{\beta}_0^r, \left(\tau_r^{(k)} \right)^{-1} \mathbb{D}^r \right) \quad \text{for all} k = 1, \ldots, K, \tag{21}$$

where $\boldsymbol{\beta}_0^r \in \mathbb{R}^{d_r^{\mathsf{F}}}$ is the prior mean and $\mathbb{D}^r$ is diagonal matrix with diagonal elements $d_{:,:}^r > 0$. We remind that for ordinal or binary variable we impose $\tau_{r^{\mathsf{OB}}} = 1$ for $r^{\mathsf{OB}} \in \mathscr{R}^{\mathsf{OB}}$, which eliminates $\tau_{r^{\mathsf{OB}}}$ from (21). Due to supposed prior independence of $\boldsymbol{\beta}^{(k),r}$ we get the following pdf:

$$p\left(\boldsymbol{\beta} \left| \boldsymbol{\tau}; \boldsymbol{\beta}_0, \mathbb{D} \right.\right) = \prod_{k=1}^{K} \prod_{r \in \mathscr{R}} \prod_{j=1}^{d_r^{\mathsf{F}}} \sqrt{\frac{\tau_r^{(k)}}{2\pi d_{j,j}^r}} \exp\left\{ -\frac{\tau_r^{(k)}}{2 d_{j,j}^r} \left(\beta_j^{(k),r} - \beta_{0,j}^r \right)^2 \right\}. \tag{22}$$

For numerical outcomes $r^{\mathsf{N}} \in \mathscr{R}^{\mathsf{Num}}$ we set precision parameters $\tau_{r^{\mathsf{N}}}^{(k)}$ to independently follow $\Gamma(a_1, a_2)$ with $a_1, a_2 > 0$ fixed hyper-parameters, which leads to

$$p(\boldsymbol{\tau} | a_1, a_2) = \prod_{k=1}^{K} \prod_{r^{\mathsf{N}} \in \mathscr{R}^{\mathsf{Num}}} \frac{a_2^{a_1}}{\Gamma(a_1)} \left(\tau_{r^{\mathsf{N}}}^{(k)} \right)^{a_1 - 1} \exp\left\{ -a_2 \tau_{r^{\mathsf{N}}}^{(k)} \right\}. \tag{23}$$

We set $\beta_{0,j}^r = 0$, $d_{j,j}^r = 10$ and $a_1 = a_2 = 1$.

For set of random effects $\boldsymbol{\mu}$ we suppose $\boldsymbol{\mu}^{(k)} | \boldsymbol{\tau}_{\mathsf{R}}^{(k)}; \boldsymbol{\mu}_0 \sim \mathsf{N}_{d^{\mathsf{R}}} \left(\boldsymbol{\mu}_0, \mathsf{diag}\left(\tau_{\mathsf{R}}^{(k)} \right)^{-1} \right)$ independently for all $k = 1, \ldots, K$, where $\boldsymbol{\mu}_0 \in \mathbb{R}^{d^{\mathsf{R}}}$ is the prior mean of $\boldsymbol{\mu}^{(k)}$ and its prior covariance matrix is diagonal with positive elements $\boldsymbol{\tau}_{\mathsf{R}}^{(k)} = \left(\tau_{\mathsf{R},1}^{(k)}, \ldots, \tau_{\mathsf{R},d^{\mathsf{R}}}^{(k)} \right)^{\top}$. These diagonal elements (viewed as another random parameters of our hierarchical model) might not be considered to be group-specific, nevertheless, we from now on work under this assumption for the sake of generality. Similarly as for elements of $\boldsymbol{\tau}$ we suppose elements of $\boldsymbol{\tau}_{\mathsf{R}}$ follow independently $\Gamma(a_3, a_4)$ with fixed hyper-parameters $a_3, a_4 > 0$, which results in probability density functions:

$$\begin{aligned} p\left(\boldsymbol{\mu} \left| \boldsymbol{\tau}_{\mathsf{R}}; \boldsymbol{\mu}_0 \right.\right) &= \prod_{k=1}^{K} \prod_{j=1}^{d^{\mathsf{R}}} \sqrt{\frac{\tau_{\mathsf{R},j}^{(k)}}{2\pi}} \exp\left\{ -\frac{\tau_{\mathsf{R},j}^{(k)}}{2} \left(\mu_j^{(k)} - \mu_{0,j}^{(k)} \right)^2 \right\}, \\ p\left(\boldsymbol{\tau}_{\mathsf{R}} \left| a_3, a_4 \right.\right) &= \prod_{k=1}^{K} \prod_{j=1}^{d^{\mathsf{R}}} \frac{a_4^{a_3}}{\Gamma(a_3)} \left(\tau_{\mathsf{R},j}^{(k)} \right)^{a_3 - 1} \exp\left\{ -a_4 \tau_{\mathsf{R},j}^{(k)} \right\}. \end{aligned} \tag{24}$$

We set $\mu_{0,j}^{(k)} = 0$ and $a_3 = a_4 = 1$.

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 $\boldsymbol{\Sigma}^{-(k)}$ to follow Wishart distribution to preserve conjugacy. More specifically, $\boldsymbol{\Sigma}^{-(k)} \left| \mathbb{Q}^{-(k)}, \nu_0 \sim \mathsf{W}_{d^{\mathsf{R}}} \left(\mathbb{Q}^{(k)}, \nu_0 \right) \right.$, where $\mathbb{Q}^{(k)}$ is (for generality purposes) group-specific scale matrices and $\nu_0 \geq d^{\mathsf{R}}$ is the number of degrees of freedom. Auxiliary positive definite random matrix $\mathbb{Q}^{-(k)}$ is distributed in very similar way: $\mathbb{Q}^{-(k)} \left| \mathbb{D}^{\mathbb{Q}}, \nu_1 \sim \mathsf{W}_{d^{\mathsf{R}}} \left(\mathbb{D}^{\mathbb{Q}}, \nu_1 \right) \right.$, where $\mathbb{D}^{\mathbb{Q}}$ is diagonal matrix with fixed diagonal elements $d_{:,:}^{\mathbb{Q}} > 0$ and ν_1 is again the number of degrees of freedom. Probability density functions of $\boldsymbol{\Sigma}^{-1} = \left\{ \boldsymbol{\Sigma}^{-(k)}, k = 1, \ldots, K \right\}$ and $\mathbb{Q}^{-1} = \left\{ \mathbb{Q}^{-(k)}, k = 1, \ldots, K \right\}$ then can be expressed as

$$\begin{aligned} p\left(\boldsymbol{\Sigma}^{-1} \left| \mathbb{Q}^{-1}; \nu_0 \right.\right) &\propto \prod_{k=1}^{K} \left| \mathbb{Q}^{-(k)} \right|^{\frac{\nu_0}{2}} \left| \boldsymbol{\Sigma}^{-(k)} \right|^{\frac{\nu_0 - d^{\mathsf{R}} - 1}{2}} \exp\left\{ -\mathsf{Tr}\left[\mathbb{Q}^{-(k)} \boldsymbol{\Sigma}^{-(k)} \right] \right\}, \\ p\left(\mathbb{Q}^{-1} \left| \mathbb{D}^{\mathbb{Q}}, \nu_1 \right.\right) &\propto \prod_{k=1}^{K} \left| \mathbb{Q}^{-(k)} \right|^{\frac{\nu_1 - d^{\mathsf{R}} - 1}{2}} \exp\left\{ -\mathsf{Tr}\left[\left(\mathbb{D}^{\mathbb{Q}} \right)^{-1} \mathbb{Q}^{-(k)} \right] \right\}. \end{aligned} \tag{25}$$

We set $\nu_0 = \nu_1 = d^{\mathsf{R}} + 1$ and $\mathbb{D}^{\mathbb{Q}} = 100 \cdot \mathbb{I}_{d^{\mathsf{R}}}$.

The set of all hyper-parameters then consists of

$$\boldsymbol{\zeta} = \left\{ \boldsymbol{\alpha}, \gamma_1^{r^{\mathsf{OB}}}, \boldsymbol{\beta}_0^r, \mathbb{D}^r, a_1, a_2, \boldsymbol{\mu}_0, a_3, a_4, \nu_0, \nu_1, \mathbb{D}^{\mathbb{Q}}, r^{\mathsf{OB}} \in \mathscr{R}^{\mathsf{OB}}, r \in \mathscr{R} \right\}.$$

4.6.2 Gibbs algorithm

For construction of Gibbs algorithm it is needed to derive full-conditioned distributions of all random parameters that enter our model $\boldsymbol{\Psi} = \left\{\boldsymbol{U}, \mathbb{Y}^{\star,\mathsf{OB}}, \boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{\tau}, \boldsymbol{\tau}_{\mathsf{R}}, \mathbb{Q}^{-1}\right\}$. By full-conditioned distribution of parameter $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ we understand distribution $\boldsymbol{\psi}|\mathbb{Y}, \mathscr{C}; \boldsymbol{\Psi}_{-\boldsymbol{\psi}}; \boldsymbol{\zeta}$, where $\boldsymbol{\Psi}_{-\boldsymbol{\psi}} = \boldsymbol{\Psi} \setminus \{\boldsymbol{\psi}\}$, with probability density function

$$p\left(\boldsymbol{\psi}\middle|\mathbb{Y}, \mathscr{C}; \boldsymbol{\Psi}_{-\boldsymbol{\psi}}; \boldsymbol{\zeta}\right) \propto \left[\prod_{i=1}^{n} p\left(\mathbb{Y}_i, U_i, \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{b}_i \middle| \mathscr{C}_i; \boldsymbol{\theta}\right)\right] \cdot p\left(\boldsymbol{\theta}, \mathbb{Q}^{-1}, \boldsymbol{\tau}_{\mathsf{R}} \middle| \boldsymbol{\zeta}\right) \tag{26}$$

viewed as a function of $\boldsymbol{\psi}$.

Derivations of all full-conditioned distributions of $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ are postponed to Appendix section A. In this section we present only the resulting Gibbs sampling algorithm using the notation declared in Appendix.

Algorithm By index $\bullet_{[m]}$ we understand value of parameter $\bullet$ generated in m-th step.

1. Take initial values $\boldsymbol{\psi}_{[0]}$ and set $m = 1$.
2. Generate m-th set of parameters $\boldsymbol{\psi}_{[m]}$:
   (a) $\boldsymbol{w}|\boldsymbol{U}; \boldsymbol{\alpha} \overset{(37)}{\sim} \mathsf{Dir}_K\left(\boldsymbol{n}(\boldsymbol{U}) + \boldsymbol{\alpha}\right)$,
   (b) $\mathsf{P}\left(U_i = k \middle| \mathbb{Y}_i^{\mathsf{N}}, \mathscr{C}_i; \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{w}\right) \overset{(38)}{=} \cdots$,
   (c) $Y_{i,j}^{\star,r^{\mathsf{OB}}} \middle| Y_{i,j}^{r^{\mathsf{OB}}} = l, \boldsymbol{\gamma} \overset{(39)}{\sim} \mathsf{TN}\left(\eta_{i,j}^{(U_i),r^{\mathsf{OB}}}, 1, \gamma_l^{r^{\mathsf{OB}}}, \gamma_{l+1}^{r^{\mathsf{OB}}}\right)$,
   (d) $\gamma_l^{r^{\mathsf{O}}} \middle| \boldsymbol{Y}^{r^{\mathsf{O}}}; \boldsymbol{Y}^{\star,r^{\mathsf{O}}} \overset{(40)}{\sim} \mathsf{Unif}\left[\max_{y \in \mathscr{Y}_{l-1}^{r^{\mathsf{O}}}} y, \min_{y \in \mathscr{Y}_l^{r^{\mathsf{O}}}} y\right], \qquad l = 1, \ldots, L^{r^{\mathsf{O}}}$,
   (e) $\tau_{r^{\mathsf{N}}}^{(k)} \middle| \boldsymbol{Y}^{r^{\mathsf{N}}}, \mathscr{C}^{r^{\mathsf{N}}}; \boldsymbol{U}, \boldsymbol{b}^{r^{\mathsf{N}}}; \boldsymbol{\beta}^{(k),r^{\mathsf{N}}}; \boldsymbol{\beta}_0^{r^{\mathsf{N}}}, \mathbb{D}^{r^{\mathsf{N}}}, a_1, a_2 \overset{(42)}{\sim} \Gamma\left(\widetilde{a}_1^{(k),r^{\mathsf{N}}}, \widetilde{a}_2^{(k),r^{\mathsf{N}}}\right)$,
   (f) $\boldsymbol{\beta}^{(k),r} \middle| \boldsymbol{Y}^r, \mathscr{C}^r; \boldsymbol{U}, \boldsymbol{b}^r; \tau_r^{(k)}; \boldsymbol{\beta}_0^r, \mathbb{D}^r \overset{(44)}{\sim} \mathsf{N}_{d_r^{\mathsf{F}}}\left(\widetilde{\boldsymbol{\beta}}^{(k),r}, \frac{\left[\left(\mathbb{X}^r_{\mathscr{N}_k(\boldsymbol{U})}\right)^\top \mathbb{X}^r_{\mathscr{N}_k(\boldsymbol{U})} + \left(\mathbb{D}^r\right)^{-1}\right]^{-1}}{\tau_r^{(k)}}\right)$,
   (g) $\tau_{\mathsf{R},j}^{(k)} \middle| \mu_j^{(k)}; \mu_{0,j}^{(k)}, a_3, a_4 \overset{(46)}{\sim} \Gamma\left(a_3 + \frac{1}{2}, a_4 + \frac{1}{2}\left(\mu_j^{(k)} - \mu_{0,j}^{(k)}\right)^2\right)$,
   (h) $\boldsymbol{\mu}^{(k)} \middle| \boldsymbol{U}, \boldsymbol{b}; \boldsymbol{\Sigma}^{-(k)}; \boldsymbol{\tau}_{\mathsf{R}}^{(k)}, \boldsymbol{\mu}_0^{(k)} \overset{(48)}{\sim} \mathsf{N}_{d^{\mathsf{R}}}\left(\widetilde{\boldsymbol{\mu}}^{(k)}, \left[n^k(\boldsymbol{U})\boldsymbol{\Sigma}^{-(k)} + \mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right)\right]^{-1}\right)$,
   (i) $\mathbb{Q}^{-(k)} \middle| \boldsymbol{\Sigma}^{-(k)}; \nu_0, \nu_1, \mathbb{D}^{\mathbb{Q}} \overset{(50)}{\sim} \mathsf{W}_{d^{\mathsf{R}}}\left(\left[\boldsymbol{\Sigma}^{-(k)} + \left(\mathbb{D}^{\mathbb{Q}}\right)^{-1}\right]^{-1}, \nu_0 + \nu_1\right)$,
   (j) $\boldsymbol{\Sigma}^{-(k)} \middle| \boldsymbol{U}, \boldsymbol{b}; \boldsymbol{\mu}^{(k)}, \mathbb{Q}^{-(k)}; \nu_0 \overset{(52)}{\sim} \mathsf{W}_{d^{\mathsf{R}}}\left(\widetilde{\mathbb{Q}}^{(k)}, n^k(\boldsymbol{U}) + \nu_0\right)$,
   (k) $\boldsymbol{b}_i \middle| \mathbb{Y}_i^{\mathsf{N}}, \mathscr{C}_i; \mathbb{Y}_i^{\star,\mathsf{OB}}, U_i; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \overset{(54)}{\sim} \mathsf{N}_{d^{\mathsf{R}}}\left(\widetilde{\boldsymbol{b}}_i, \left[\widetilde{\mathbb{Z}}_i^\top \widetilde{\mathbb{Z}}_i + \boldsymbol{\Sigma}^{-(k)}\right]^{-1}\right)$,

   where in each step we use the latest generated values of parameters, i.e. either elements of previously generated $\boldsymbol{\psi}_{[m-1]}$ or currently generated $\boldsymbol{\psi}_{[m]}$.
3. If $m < M$ then increase $m := m + 1$ and repeat step 2., otherwise end the algorithm.
4. Output: M generated states $\boldsymbol{\psi}_{[1]}, \ldots, \boldsymbol{\psi}_{[M]}$.

Generated values $\boldsymbol{\psi}_{[1]}, \ldots, \boldsymbol{\psi}_{[M]}$ are then used for estimation of marginal distributions of individual parameters. Usual methods as cutting first B values (burn-in period) that serve to get to posterior distribution and thinning the rest of the chain to lower the autocorrelation were applied.

For example, the final classification of subject i into one of K latent groups can be based on generated values $U_{i,[1]}, \ldots, U_{i,[M]}$. When cutting the burn-in period (of length $B < M$) and thinning (of order t) are applied then subject i can be classified into the most common value among the rest of generated U_i:

$$\widehat{U}_i = \mod\left\{U_{i,[B+m*t+1]}, m \in \left\{0, 1, \ldots, \left\lfloor \frac{M-B-1}{t} \right\rfloor\right\}\right\}. \tag{27}$$

For practical reasons it also makes sense to actually classify subject into this group only if the corresponding proportion is higher than given threshold, if not, the subject remains unclassified to avoid misclassification.

Another way of classification is presented in the next section, which also covers the situation of newly observed subject.

4.6.3 Label switching

One of the possible disadvantages of MCMC approach is the fact that the latent clusters $1,\dots,K$ do not have fixed meaning. The resulting estimates of class-specific parameters could be arbitrarily permuted while preserving the final model. There is a danger of possible change of the cluster interpretation during a single chain called *label switching*. This problem could be dealt within post-processing phase by following recommendations of Stephens (2000). As we consider low values of K we adopt an algorithm that examines all $K!$ possibilities for each iteration and chooses the one minimizing a given criterion. However, during our experiments no need of permutation has occurred. This may be given by the complexity of our model that prevents any switching of the interpretation during our Gibbs sampling procedure.

4.7 Classification probabilities calculation

Let us consider subject with outcomes $\mathbb{Y}_{\mathsf{new}}$ and covariate values $\mathscr{C}_{\mathsf{new}}$, our task is to classify this subject into one of K latent groups. In Section **??** we described mixture probability density function (8) and (**??**) of all outcomes given covariates for one subject. Derived classification probability $u_{i,k}$ is a function that depends on parameters of interest $\boldsymbol{\theta}$ but is free of latent parameters such as $\mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}}$, $\boldsymbol{b}_{\mathsf{new}}$ or $\boldsymbol{\tau}_{\mathsf{R}}$, $\mathbb{Q}^{-1}$. Such a parametric function can be evaluated for each generated state $\boldsymbol{\theta}_{[m]}$ to obtain a chain of posterior classification probabilities. Unfortunately, each evaluation of such function requires integration with respect to random effects $\boldsymbol{b}_{\mathsf{new}}$ and latent numeric outcomes $\mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}}$ described in (8):

$$\begin{aligned}
&\int\int p\left(\mathbb{Y}_{\mathsf{new}}^{\mathsf{OB}}\middle|\,\mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}};\boldsymbol{\gamma}\right)\cdot p\left(\mathbb{Y}_{\mathsf{new}}^{\mathsf{N}},\mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}}\middle|\,\mathscr{C}_{\mathsf{new}},\boldsymbol{b}_{\mathsf{new}};\boldsymbol{\beta},\boldsymbol{\tau}\right)\cdot p\left(\boldsymbol{b}_{\mathsf{new}}|\,\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\right)\mathrm{d}\boldsymbol{b}_{\mathsf{new}}\,\mathrm{d}\mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}} = \\
&\quad=\int p\left(\mathbb{Y}_{\mathsf{new}}^{\mathsf{OB}}\middle|\,\mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}};\boldsymbol{\gamma}\right)\cdot\underbrace{\left[\int p\left(\mathbb{Y}_{\mathsf{new}}^{\mathsf{N}},\mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}}\middle|\,\mathscr{C}_{\mathsf{new}},\boldsymbol{b}_{\mathsf{new}};\boldsymbol{\beta},\boldsymbol{\tau}\right)\cdot p\left(\boldsymbol{b}_{\mathsf{new}}|\,\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\right)\mathrm{d}\boldsymbol{b}_{\mathsf{new}}\right]}_{p\left(\mathbb{Y}_{\mathsf{new}}^{\mathsf{N}},\mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}}\middle|\mathscr{C}_{\mathsf{new}};\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\right)}\mathrm{d}\mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}} \qquad (28)
\end{aligned}$$

4.7.1 Integration with respect to random effects $\mathbf{b}_{\mathsf{new}}$

Let us first perform the integration of random effects $\boldsymbol{b}_{\mathsf{new}}$ from (11) to obtain distribution of numeric variables unconditioned by random effects. This issue is commonly faced and solved in LME models by realizing that under normality assumption of both numeric outcomes and random effects the unconditioned distribution of outcomes is also normal. Let us stack all outcomes into one long vector:

$$\boldsymbol{Y}_{\mathsf{new}}=\begin{pmatrix}\vdots\\ \boldsymbol{Y}_{\mathsf{new}}^{r^{\mathsf{N}}}\\ \vdots\\ \boldsymbol{Y}_{\mathsf{new}}^{\star,r^{\mathsf{OB}}}\\ \vdots\end{pmatrix},\quad \begin{matrix}r^{\mathsf{N}}\in\mathscr{R}^{\mathsf{Num}},\\ r^{\mathsf{OB}}\in\mathscr{R}^{\mathsf{OB}}\end{matrix}\quad\text{and stack regressors into diagonal matrices}$$

$$\mathbb{X}_{\mathsf{new}}=\begin{pmatrix}\ddots & & & \\ & \mathbb{X}_{\mathsf{new}}^{r^{\mathsf{N}}} & & \\ & & \ddots & \\ & & & \mathbb{X}_{\mathsf{new}}^{r^{\mathsf{OB}}} & \\ & & & & \ddots\end{pmatrix},\ \mathbb{Z}_{\mathsf{new}}=\begin{pmatrix}\ddots & & & \\ & \mathbb{Z}_{\mathsf{new}}^{r^{\mathsf{N}}} & & \\ & & \ddots & \\ & & & \mathbb{Z}_{\mathsf{new}}^{r^{\mathsf{OB}}} & \\ & & & & \ddots\end{pmatrix},\quad \begin{matrix}r^{\mathsf{N}}\in\mathscr{R}^{\mathsf{Num}},\\ r^{\mathsf{OB}}\in\mathscr{R}^{\mathsf{OB}}.\end{matrix}$$

In this section let $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\tau}$ stand for only one representative of the parameter among K groups (similarly as in opening Section **??**). Moreover, let $\boldsymbol{\beta}$ and $\boldsymbol{b}_{\mathsf{new}}$ be stacked vectors of fixed and random effects of all outcomes,

i. e.

$$\boldsymbol{\beta} = \begin{pmatrix} \vdots \\ \boldsymbol{\beta}^{r^{\mathsf{N}}} \\ \vdots \\ \boldsymbol{\beta}^{r^{\mathsf{OB}}} \\ \vdots \end{pmatrix}, \qquad \boldsymbol{b}_{\mathsf{new}} = \begin{pmatrix} \vdots \\ \boldsymbol{b}_{\mathsf{new}}^{r^{\mathsf{N}}} \\ \vdots \\ \boldsymbol{b}_{\mathsf{new}}^{r^{\mathsf{OB}}} \\ \vdots \end{pmatrix}, \qquad \begin{matrix} r^{\mathsf{N}} \in \mathscr{R}^{\mathsf{Num}}, \\ r^{\mathsf{OB}} \in \mathscr{R}^{\mathsf{OB}}. \end{matrix}$$

Adding dimensions $d = n_{\mathsf{new}} |\mathscr{R}|$, $d^{\mathsf{N}} = n_{\mathsf{new}} \left|\mathscr{R}^{\mathsf{Num}}\right|$, $d^{\mathsf{OB}} = n_{\mathsf{new}} \left|\mathscr{R}^{\mathsf{OB}}\right|$

$$\boldsymbol{Y}_{\mathsf{new}} | \boldsymbol{b}_{\mathsf{new}} \sim \mathsf{N}_d \left(\mathbb{X}_{\mathsf{new}} \boldsymbol{\beta} + \mathbb{Z}_{\mathsf{new}} \boldsymbol{b}_{\mathsf{new}}, \mathbb{T} = \mathsf{diag} \left\{ \tau_{r^{\mathsf{N}}}^{-1} \mathbb{I}_{n_{\mathsf{new}}}, r^{\mathsf{N}} \in \mathscr{R}^{\mathsf{Num}}, \mathbb{I}_{d^{\mathsf{OB}}} \right\} \right)$$

Now we can use notoriously known formulas to obtain unconditioned mean and variance matrix:

$$\begin{aligned} \mathsf{E} \boldsymbol{Y}_{\mathsf{new}} &= \mathsf{E} \left(\mathsf{E} \left[\boldsymbol{Y}_{\mathsf{new}} | \boldsymbol{b}_{\mathsf{new}} \right] \right) = \mathbb{X}_{\mathsf{new}} \boldsymbol{\beta} + \mathbb{Z}_{\mathsf{new}} \boldsymbol{\mu}, \\ \mathsf{var} \boldsymbol{Y}_{\mathsf{new}} &= \mathsf{E} \left(\mathsf{var} \left[\boldsymbol{Y}_{\mathsf{new}} | \boldsymbol{b}_{\mathsf{new}} \right] \right) + \mathsf{var} \left(\mathsf{E} \left[\boldsymbol{Y}_{\mathsf{new}} | \boldsymbol{b}_{\mathsf{new}} \right] \right) = \mathbb{T} + \mathbb{Z}^{\top} \boldsymbol{\Sigma} \mathbb{Z} =: \mathbb{V}, \end{aligned}$$

which results in

$$\boldsymbol{Y}_{\mathsf{new}} \sim \mathsf{N}_d \left(\mathbb{X}_{\mathsf{new}} \boldsymbol{\beta} + \mathbb{Z}_{\mathsf{new}} \boldsymbol{\mu}, \mathbb{T} + \mathbb{Z}^{\top} \boldsymbol{\Sigma} \mathbb{Z} \right), \tag{29}$$

which can be distribution with completely general covariance structure due to our generality of $\boldsymbol{\Sigma}$. Thus, dependencies among outcomes are captured by this model.

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

It remains to perform the following integration:

$$\int p \left(\mathbb{Y}_{\mathsf{new}}^{\mathsf{OB}} \middle| \mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}}; \boldsymbol{\gamma} \right) \cdot p \left(\mathbb{Y}_{\mathsf{new}}^{\mathsf{N}}, \mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}} \middle| \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \right) \mathrm{d} \mathbb{Y}_{\mathsf{new}}^{\star,\mathsf{OB}},$$

which is in fact integration of multivariate normal density within the bounds given by thresholds $\boldsymbol{\gamma}$ and observed ordinal and binary outcomes. Only the truly numerical part survives this integration intact, therefore, we aim to separate it from the rest.

Let us divide vector $\boldsymbol{Y}_{\mathsf{new}}$ on the truly numerical part and the vector of latent numeric variables for ordinal and binary outcomes:

$$\begin{pmatrix} \boldsymbol{Y}_{\mathsf{new}}^{\mathsf{N}} \\ \boldsymbol{Y}_{\mathsf{new}}^{\star,\mathsf{OB}} \end{pmatrix} \sim \mathsf{N}_{d^{\mathsf{N}} + d^{\mathsf{OB}}} \left(\begin{pmatrix} \boldsymbol{\eta}^{\mathsf{N}} \\ \boldsymbol{\eta}^{\mathsf{OB}} \end{pmatrix} = \begin{pmatrix} \mathbb{X}_{\mathsf{new}}^{\mathsf{N}} \boldsymbol{\beta}^{\mathsf{N}} + \mathbb{Z}_{\mathsf{new}}^{\mathsf{N}} \boldsymbol{\mu}^{\mathsf{N}} \\ \mathbb{X}_{\mathsf{new}}^{\mathsf{OB}} \boldsymbol{\beta}^{\mathsf{OB}} + \mathbb{Z}_{\mathsf{new}}^{\mathsf{OB}} \boldsymbol{\mu}^{\mathsf{OB}} \end{pmatrix}, \begin{pmatrix} \mathbb{V}_{\mathsf{N}} & \mathbb{V}_{\mathsf{NOB}} \\ \mathbb{V}_{\mathsf{OBN}} & \mathbb{V}_{\mathsf{OB}} \end{pmatrix} \right). \tag{30}$$

Since the latent part is about to be integrated out we exploit properties of multivariate normal distribution to obtain

$$\begin{aligned} \boldsymbol{Y}_{\mathsf{new}}^{\mathsf{N}} &\sim \mathsf{N}_{d^{\mathsf{N}}} \left(\boldsymbol{\eta}^{\mathsf{N}}, \mathbb{V}_{\mathsf{N}} \right), \\ \boldsymbol{Y}_{\mathsf{new}}^{\star,\mathsf{OB}} \middle| \boldsymbol{Y}_{\mathsf{new}}^{\mathsf{N}} = \boldsymbol{y}_{\mathsf{new}}^{\mathsf{N}} &\sim \mathsf{N}_{d^{\mathsf{OB}}} \left(\boldsymbol{\eta}'_{\mathsf{OB}}, \mathbb{V}'_{\mathsf{OB}} \right), \end{aligned}$$

where $\boldsymbol{\eta}'_{\mathsf{OB}} = \boldsymbol{\eta}^{\mathsf{OB}} + \mathbb{V}_{\mathsf{OBN}} \mathbb{V}_{\mathsf{N}}^{-1} \left(\boldsymbol{y}_{\mathsf{new}}^{\mathsf{N}} - \boldsymbol{\eta}^{\mathsf{N}} \right)$ and $\mathbb{V}'_{\mathsf{OB}} = \mathbb{V}_{\mathsf{OB}} - \mathbb{V}_{\mathsf{OBN}} \mathbb{V}_{\mathsf{N}}^{-1} \mathbb{V}_{\mathsf{NOB}}$ and corresponding probability distribution function decomposes into the product

$$\begin{aligned} p \left(\boldsymbol{Y}_{\mathsf{new}}^{\mathsf{N}}, \boldsymbol{Y}_{\mathsf{new}}^{\star,\mathsf{OB}} \middle| \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \right) = \\ = p \left(\boldsymbol{Y}_{\mathsf{new}}^{\star,\mathsf{OB}} \middle| \boldsymbol{Y}_{\mathsf{new}}^{\mathsf{N}}, \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \right) \cdot p \left(\boldsymbol{Y}_{\mathsf{new}}^{\mathsf{N}} \middle| \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \right), \end{aligned}$$

second part of which is invariant of $\boldsymbol{Y}_{\mathsf{new}}^{\star,\mathsf{OB}}$ and, therefore, avoids integration:

$$p \left(\boldsymbol{Y}_{\mathsf{new}}^{\mathsf{N}} \middle| \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \right) \cdot \int \underbrace{p \left(\boldsymbol{Y}_{\mathsf{new}}^{\mathsf{OB}} \middle| \boldsymbol{Y}_{\mathsf{new}}^{\star,\mathsf{OB}}; \boldsymbol{\gamma} \right)}_{\text{thresholding (3)}} \cdot \underbrace{p \left(\boldsymbol{Y}_{\mathsf{new}}^{\star,\mathsf{OB}} \middle| \boldsymbol{Y}_{\mathsf{new}}^{\mathsf{N}}, \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \right)}_{\text{pdf of } \mathsf{N}_{d^{\mathsf{OB}}} \left(\boldsymbol{\eta}'_{\mathsf{OB}}, \mathbb{V}'_{\mathsf{OB}} \right)} \mathrm{d} \boldsymbol{Y}_{\mathsf{new}}^{\star,\mathsf{OB}}.$$

I zde je třeba se rozhodnout, jak s tím tlustým $\boldsymbol{Y}$ či $\mathbb{Y}$.

It remains to integrate the product of two functions, first one of which just declares lower and upper integration bounds and the second one is the probability density function of multivariate normal distribution $\mathsf{N}_{d^{\mathsf{OB}}}\left(\boldsymbol{\eta}'_{\mathsf{OB}}, \mathbb{V}'_{\mathsf{OB}}\right)$. For each individual categorical outcome $r^{\mathsf{OB}} \in \mathscr{R}^{\mathsf{OB}}$ and observation $j \in \{1, \ldots, n_{\mathsf{new}}\}$ the value $y^{r^{\mathsf{OB}}}_{\mathsf{new},j} = l$ determines an interval given by the corresponding pair of γ parameters, see (3):

$$Y^{r^{\mathsf{OB}}}_{\mathsf{new},j} = l \quad \Longrightarrow \quad Y^{\star,r^{\mathsf{OB}}}_{\mathsf{new},j} \in \left(\gamma^{r^{\mathsf{OB}}}_{l}, \gamma^{r^{\mathsf{OB}}}_{l+1}\right] =: \left(e^{r^{\mathsf{OB}}}_{j}, f^{r^{\mathsf{OB}}}_{j}\right].$$

If we denote the resulting Cartesian product of these intervals as $\square\left(\boldsymbol{Y}^{\mathsf{OB}}_{\mathsf{new}}\right) \subset \mathbb{R}^{d^{\mathsf{OB}}}$, i.e.

$$\square\left(\boldsymbol{Y}^{\mathsf{OB}}_{\mathsf{new}}\right) = (\boldsymbol{e}, \boldsymbol{f}] = \bigotimes_{r^{\mathsf{OB}} \in \mathscr{R}^{\mathsf{OB}}} \bigotimes_{j=1}^{n_{\mathsf{new}}} \left(e^{r^{\mathsf{OB}}}_{j}, f^{r^{\mathsf{OB}}}_{j}\right] = \bigotimes_{r^{\mathsf{OB}} \in \mathscr{R}^{\mathsf{OB}}} \bigotimes_{j=1}^{n_{\mathsf{new}}} \left(\gamma^{r^{\mathsf{OB}}}_{Y^{\star,r^{\mathsf{OB}}}_{\mathsf{new},j}}, \gamma^{r^{\mathsf{OB}}}_{Y^{\star,r^{\mathsf{OB}}}_{\mathsf{new},j}+1}\right],$$

then the remaining integral can be written in the form

$$I(\square) = \int\limits_{\square\left(\boldsymbol{Y}^{\mathsf{OB}}_{\mathsf{new}}\right)} p\left(\boldsymbol{Y}^{\star,\mathsf{OB}}_{\mathsf{new}} \middle| \boldsymbol{Y}^{\mathsf{N}}_{\mathsf{new}}, \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\right) \mathrm{d}\boldsymbol{Y}^{\star,\mathsf{OB}}_{\mathsf{new}} = \int\limits_{\boldsymbol{e}}^{\boldsymbol{f}} p\left(\boldsymbol{y} \middle| \boldsymbol{\eta}'_{\mathsf{OB}}, \mathbb{V}'_{\mathsf{OB}}\right) \mathrm{d}\boldsymbol{y}, \tag{31}$$

where $p\left(\boldsymbol{y} \middle| \boldsymbol{\eta}'_{\mathsf{OB}}, \mathbb{V}'_{\mathsf{OB}}\right)$ denotes the probability density function of multivariate normal distribution $\mathsf{N}_{d^{\mathsf{OB}}}\left(\boldsymbol{\eta}'_{\mathsf{OB}}, \mathbb{V}'_{\mathsf{OB}}\right)$.

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

Altogether, the probability of being classified into group k is proportional to

$$\begin{aligned} u_{\mathsf{new},k}(\boldsymbol{\theta}) &\propto w_k \cdot p\left(\boldsymbol{Y}^{\mathsf{N}}_{\mathsf{new}} \middle| \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{-(k)}\right) \cdot I_k(\square), \\ &= \frac{w_k \cdot p\left(\boldsymbol{Y}^{\mathsf{N}}_{\mathsf{new}} \middle| \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{-(k)}\right) \cdot I_k(\square)}{\sum\limits_{k'=1}^{K} w_{k'} \cdot p\left(\boldsymbol{Y}^{\mathsf{N}}_{\mathsf{new}} \middle| \mathscr{C}_{\mathsf{new}}; \boldsymbol{\beta}^{(k')}, \boldsymbol{\tau}^{(k')}, \boldsymbol{\mu}^{(k')}, \boldsymbol{\Sigma}^{-(k')}\right) \cdot I_{k'}(\square)} \end{aligned} \tag{32}$$

where I_k is the integral (13) computed with set of parameters $\boldsymbol{\theta}^{(k)}$ belonging to group k. These classification probabilities are then computed for all (or just a subset due to burn-in period and thinning) generated states $\boldsymbol{\theta}_{[1]}, \ldots, \boldsymbol{\theta}_{[M]}$ and can be used to describe the posterior distribution. Subjects then can be classified to the class with convincingly high value of posterior mean estimator $\widehat{u}_{\mathsf{new},k}$. The subject might remain unclassified, if neither of the classes seems to dominate others.

4.7.3 Deviance

Throughout the paper we treated the number of latent classes K fixed as to be selected by statistician in advance. In most circumstances, however, there is no prior knowledge of the suitable value of K to be used. Usual practise in this situation is to try several values of K and choose the one that optimizes one of the known criteria that measure the goodness of fit to the data and penalizes the number of estimated unknown parameters $\boldsymbol{\theta}$. We present here a deviance criterion that is solely based on the goodness of fit measured by the log-likelihood function.

Deviance is defined as

$$D(\boldsymbol{\theta}; \mathbb{Y}, \mathscr{C}) = -2 \log p(\mathbb{Y} | \mathscr{C}; \boldsymbol{\theta}) = -2 \sum_{i=1}^{n} \log p(\mathbb{Y}_i | \mathscr{C}_i; \boldsymbol{\theta}). \tag{33}$$

Contribution of individual i to the deviance can be expresses as

$$-2 \log p(\mathbb{Y}_i | \mathscr{C}_i; \boldsymbol{\theta}) = -2 \log \left[\sum_{k=1}^{K} \int \int p\left(\mathbb{Y}_i, \mathbb{Y}^{\star,\mathsf{OB}}_i, \boldsymbol{b}_i, U_i = k \middle| \mathscr{C}_i; \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{b}_i \, \mathrm{d}\mathbb{Y}^{\star,\mathsf{OB}}_i \right] \tag{34}$$

which includes the integration (11) which has been described above. Once we use analogous notation we can simply write

$$D(\boldsymbol{\theta};\mathbb{Y},\mathscr{C}) = -2\sum_{i=1}^{n}\log\left[\sum_{k=1}^{K} w_k \cdot p\left(\mathbb{Y}_i^{\mathsf{N}}\middle|\mathscr{C}_i;\boldsymbol{\beta}^{(k)},\boldsymbol{\tau}^{(k)},\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{-(k)}\right)\cdot I_k\left(\square\left(\mathbb{Y}_i^{\mathsf{OB}}\right)\right)\right] \tag{35}$$

where the denominator of (14) was inserted into the logarithm. In order to compute the deviance (and all related criteria) we need to basically compute classification probabilities for every individual. Deviance is, as a matter of fact, a parametric function of $\boldsymbol{\theta}$ and has, therefore, its posterior distribution, which can be estimated by calculation of deviance parametric function on a sample $\boldsymbol{\theta}_{[1]},\ldots,\boldsymbol{\theta}_{[M]}$. As the computation of classification probabilities is time-consuming we resort to thinning.

5 Simulation

In this section we design and perform a simulation study in order to demonstrate the functionality of our proposed model. 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 (numeric, binary and ordinal) is represented by just 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 corresponds to the same amount of observations per household available in the EU-SILC data.

The part of the predictor which is common to all types of variables is of the form

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

Then we suppose that each subject has his set of times $0 < t_{i,1} < t_{i,2} < t_{i,3} < t_{i,4} < 1$ which were generated as an ordered sample from uniform distribution over interval $(0,1)$. We suppose the linear parametrization of time, which is, however, different subject to 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 3 (cases 1. and 2.) or 6 (case 3.). Its variance matrix $\boldsymbol{\Sigma}$ was adequately chosen matrix of non-diagonal form, more details can be found in supplement materials.

Another 3 types of scenarios arise from the types of differences among $K = 2$ or $K = 3$ latent groups:

a) (d=intercept) - only in the intercept term - $\beta_0^{(k)}$ (case 2.) and $\boldsymbol{\mu}_0^{(k)}$ (case 1. and 3.) are class-specific, but slope parameters β_1 and $\boldsymbol{\mu}_1$ are not,
b) (d=slope) - only in the slope - $\beta_1^{(k)}$ (case 1.) and $\boldsymbol{\mu}_1^{(k)}$ (case 2. and 3.) are class-specific, but intercept terms β_0 and μ_0 are not,
c) (d=both) - both in the intercept term and the slope - $\beta_0^{(k)}$, $\beta_1^{(k)}$, $\boldsymbol{\mu}_0^{(k)}$ and $\boldsymbol{\mu}_1^{(k)}$ are class-specific.

These three types of differences are combined with the three types of random effects structure creating 9 different scenarios which are examined for $K = 2,3$ and different sample sizes n. The values of intercept and slope for each of the 9 scenarios were chosen in different ways to obtain clusters distinguishable by eye (see Figure 2 for the case $K = 3$), the true values of intercept and slope parameters can be found in Tables S**??** and S**??** in supplement materials.

The group allocation indicator U_i was always generated from uniform distribution which results in clusters of comparable sizes. All (latent) numeric outcomes were sampled with unit variance $\tau = 1$. Binary variable was obtained by threshold $\gamma_1^{\mathsf{B}} = 1$ and the ordinal one 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 resulting estimators. Estimation is based on $M = 10\,000$ sampled states from posterior distribution. To spare computation time, thinning of order 10 was applied for classification probabilities leading to 1 000 calculated probabilities per subject. We are also interested in overall probability that subject from k-th cluster is correctly classified to cluster k. For this reason we compute for each k arithmetic mean $\overline{p}_k$ of our estimates $\widehat{\overline{U}}_{i,k}$ of posterior probabilities of belonging to cluster k across all cluster members:

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

These probabilities are also estimated dynamically, i.e. using only limited amount of information, see 4.4.

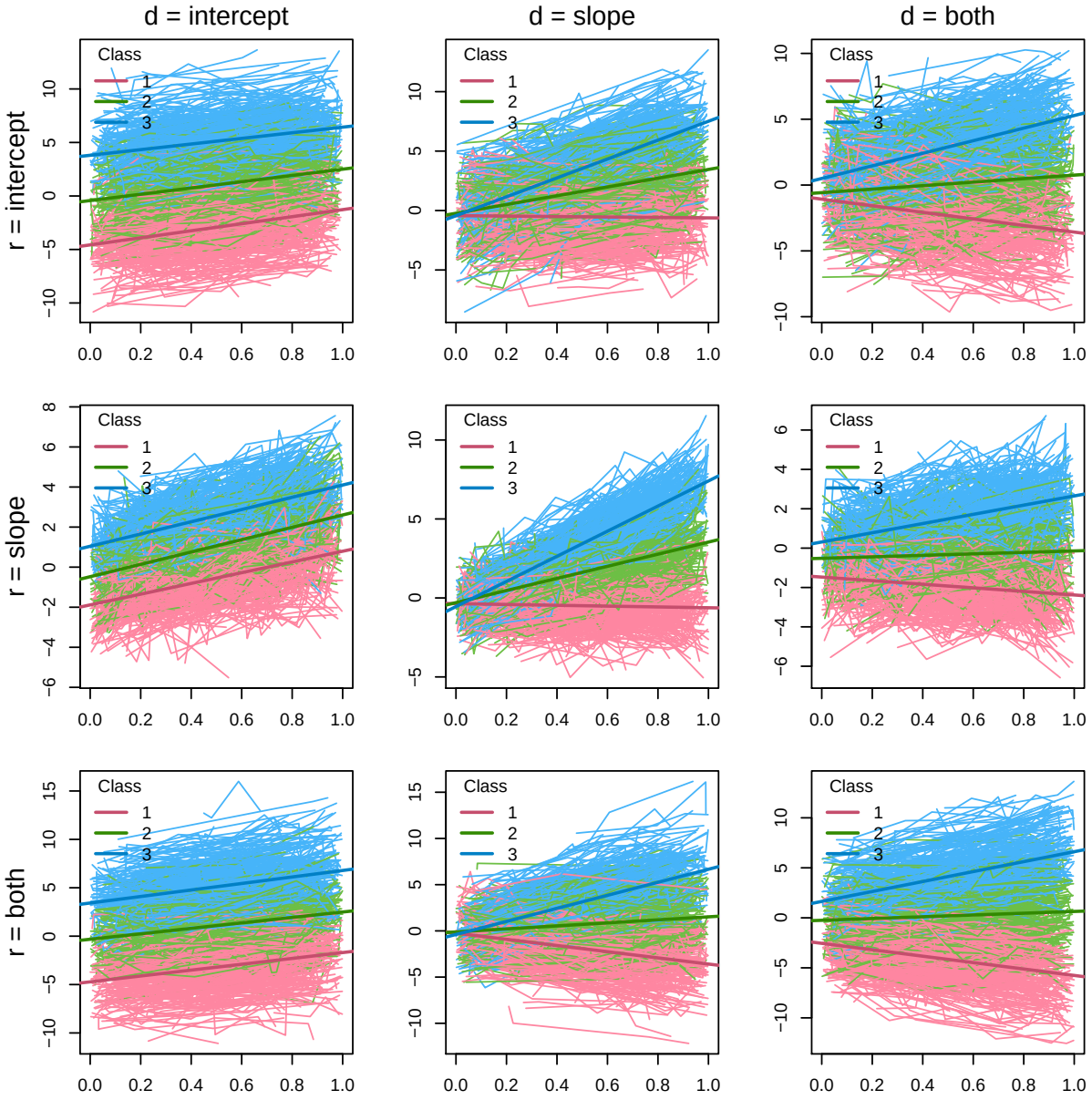


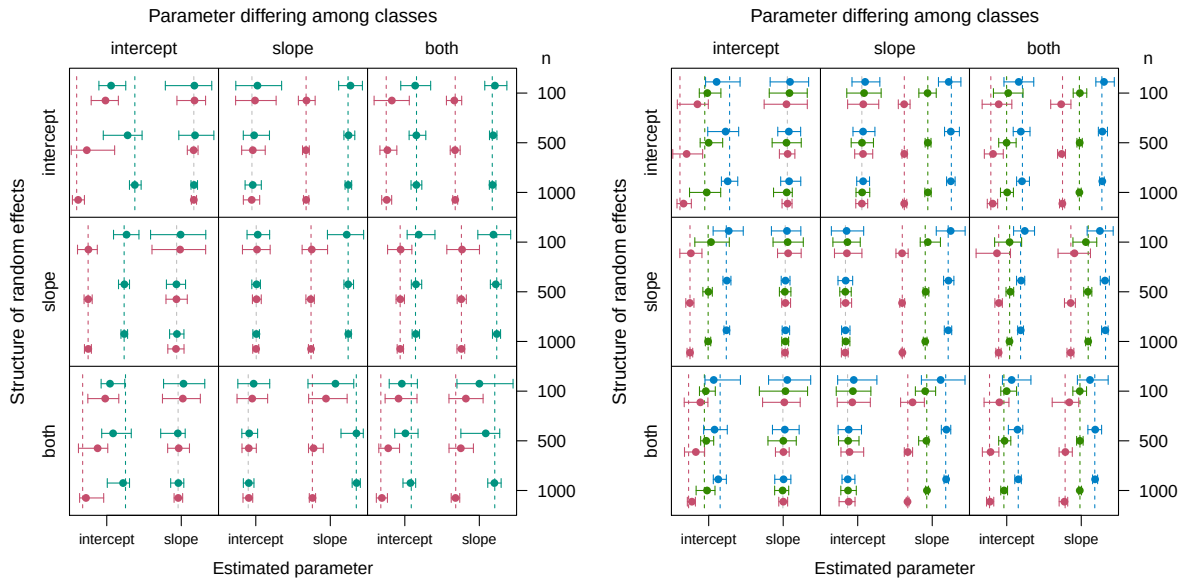
Fig. 2: The samples of a numeric outcome distinguishing different scenario types (row difference, column structure of random effects) when $K = 3$ latent groups are supposed.

5.2 Results

Let us first focus on the estimation properties of the proposed modelling procedure, which is shown by Figure 3. Colours distinguish the estimates in different classes (clusters?) and the corresponding true values of intercept and slope parameters are captured by dashed lines. Grey colour depicts the true value shared by all classes. Each segment represents 2.5% and 97.5% quantiles of 200 times replicated estimator and full circle represents corresponding mean. Figure 3 provides estimates of parameters belonging to ordinal outcome only, plots for numeric and binary are postponed to supplement materials.

Figure 3 demonstrates that the proposed procedure is capable of quite precise estimation of parameters despite the latent modelling and thresholding concept. In most cases it successfully discovers the difference among classes as intervals of different colours tend not to overlap each other. There is apparent a decreasing trend in standard deviation as n increases which is disrupted only when corresponding estimate does not reach the true value. This phenomenon occurs mostly in the estimation of intercept term when it is considered to be random and different among clusters at the same time. Such a behaviour can be spotted also for the class-specific slope term when both intercept and slope term are random effects. In these situations, the estimates are shrunk towards a mean of the true values. This might be a result of combination of incapability of discrimination between classes for low value of n and the fact that LME usually tend to shrink random effects to zero. In the case of $K = 3$, this effect does not fully vanish even for $n = 1000$, see the row *both* and column *intercept*. However, it seems that the large number of subjects n can overcome this issue, which we rely on in the further real data analysis.

As the ability to discover differences among classes has been verified we now proceed to examine the classification abilities. Table 1 contains the percentages of correctly classified subjects (using the HPD interval rule) among all n of them across 200 replications. This percentage differs scenario by scenario as the random structures and differences among classes interact in different ways leading to diverse success rates. For example, the case with 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 2. 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, much larger percentage of subjects is kept unclassified in such cases. With increasing n the percentage of unclassified subjects rapidly decreases and converts mainly into correctly classified category. 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 declines the ability to distinguish among classes, revisit Figure 2 for illustration.



(a) $K = 2$. (b) $K = 3$.

Fig. 3: 95% quantile bounds and means for the intercept and slope parameters for ordinal outcome under different random effects structures and differences between classes. True values of the parameters are depicted by dashed lines (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	$K=2$ Correct [%]	$K=2$ Uncl. [%]	$K=2$ Miscl. [%]	$K=3$ Correct [%]	$K=3$ Uncl. [%]	$K=3$ 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.

Now we discuss the classification properties when limited amount of information is given. Dynamically calculated probabilities, i.e. using just first $j \in \{1,\dots,n_i\}$ observations (see more in 4.4), are shown in Figure 4. It shows mean and quantile bounds of 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 in the random intercept term only, 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. That results in rejecting the low and the extremely large slope values of other classes, and therefore increasing the probability of classification towards the true middle class. It also improves with increasing number of subjects n since the classes are then better distinguished.

The simulation study was conducted on cluster consisting of CPU units: Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz, 64GB RAM. The average computation time of generating a chain of $M = 10\,000$ followed by much more demanding computation of 1 000 classification probabilities for all n subjects could not fit under an hour even for the lowest values of $n = 100$ and $K = 2$ (around 80 minutes). The number of calls of `pmvnorm` seem to influence the computational time the most. As expected, in the case of $n = 500$ or $n = 1\,000$ the procedure requires 5 or 10-times more time than if $n = 100$. Similarly, addition of another cluster increases the computational time as well due to one additional call of `pmvnorm` function per subject and set of $\boldsymbol{\theta}$ parameters to calculate corresponding classification probabilities. The most challenging combination of $n = 1\,000$ and $K = 3$ took around 1 200 minutes (20 hours).

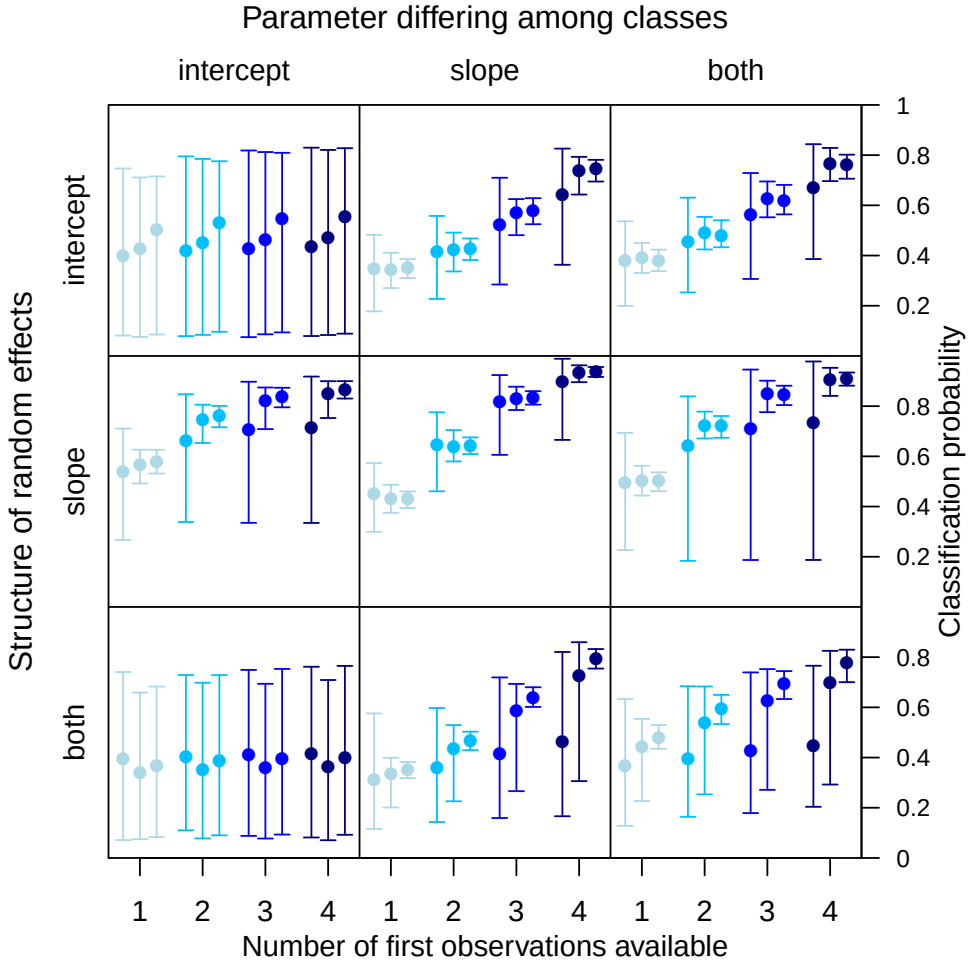


Fig. 4: Subjects of class 2 when $K = 3$. The mean and 2.5% and 97.5% quantile of mean classification probabilities $\overline{p}_2$ towards the true class calculated dynamically using only first $j \in \{1,2,3,4\}$ observations under several random effects structure and difference among $K = 3$ classes.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

6 Application to EU-SILC data

We will now apply the proposed methodology to find temporal patterns for households in the EU-SILC database in the Czech Republic as outlined in the introduction. The chosen time period 2005 – 2016 covers the economical crisis and we expect it to heavily impact the budget of households leading to different ways of coping with the crisis. From those who were not affected and prosper on, 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 we aim to discover several different patterns in their evolution while modelling them jointly.

6.1 Data description

First, we need to delve into the data gathering mechanism which is crucial for appropriate interpretation of results. The EU-SILC longitudinal study follows a rotational design - rotating a part of the sample from one year to the next and retaining the other part unchanged. 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 leave from the study, households were followed no longer than 4 years. Since the primary focus is on the evolution part we use for the analysis only the households that were interviewed indeed $n_i = 4$ 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 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 follows heavily skewed distribution. Therefore, we rather work with its logarithmic transformation which suits our LME assumptions much better. Binary outcome refers to the affordability and to the actual meaning 'ability to pay' regardless of whether the household wants it. Ordinal outcome 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 completely independent random variables.

Data contains 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 weighted family size which is a sum of weights of each of the household members. Adult person in the role of the head of the family has weight 1 and others have either weight 0.5 or 0.3 depending on whether they are older or younger than 14, respectively.

6.2 Model structure

Since the three outcomes are strongly related we cannot model them separately using independent models. Our proposed joint model capable of capturing the relationships is more than suitable for modelling of these outcomes.

We suppose the LME model of the same structure of both fixed and random component for numeric outcome and latent numeric counterparts of binary and ordinal outcomes. Being aware of possible change in the evolution of these outcomes within the time period 2005 – 2016 we decided to parametrize the effect of time by B-spline of order 3 with knots in the years 2005, 2008, 2010 and 2017, which leads to six β parameters including the intercept. This fixed part of the model is extended by the weighted family size as an additional regressor. 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 different level than others. The model formula for j-th observation of i-th household at time $t_{i,j}$ is then:

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

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, $S_{i,j}$ is the weighted family size 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.

We suppose that the hidden latent groups of households with different evolution patterns share the same covariance properties, meaning that parameters $\boldsymbol{\Sigma}$ and τ are supposed to be common for all classes. Hence, the relationships among outcomes are kept the same in all classes. However, we suppose them to differ in the $\boldsymbol{\beta}^{(k)}$ coefficients, which mainly describes the different evolution patterns captured by splines. The revealed clusters will then be characterized according to the shape and level of resulting curve.

6.3 Results

Contrary to the previous simulation study we had to be more careful when sampling from the posterior distribution. Initial values of all the unknown parameters and latent variables were randomly generated to obtain different starting points for sampled chains. The progress in each of the model parameters was visually monitored every 10 000 steps in order to suggest 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 parameters γ due to almost negligible steps. A final chain of length $M = 10\,000$ 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 classification probabilities and deviance we again thinned the chain by 10 to spare the computation time.

Following Section 4.7.3, we applied this methodology under several different choices of the number of hidden clusters K and examined the posterior distribution of deviance in hope of selecting the most suitable one. Ideally, we would search for the K such that the decrease of deviance becomes negligible. Although, some improvement in decrement of deviance is visible in Figure 5, we can also notice that the solution for $K = 4$ surprisingly achieved lower deviance than the one for $K = 5$. Higher choice of K led to even lower deviance, however, first signs of overfitting appeared with it. Small groups of households of extraordinary and very specific behaviour emerged.

The solution for $K = 5$ already contains an example of such rare clusters. Figure 6 contains estimated splines curves for 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) that seems to be followed by majority of households even for higher K. With $K = 3$ there appears new violet cluster that follows the same shape of the general curve but on much higher level. Hence, it represents about 5–9 % of households having 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 cluster of U-shaped trend (blue curve). For $K \geq 5$ such cluster appears again accompanied with a golden cluster that behaves reversely. This is why this solution should be rather viewed as an extension of $K = 3$ 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 seem to be the reason why this solution defeats $K = 5$ in terms of deviance. Unfortunately, it brings more confusion

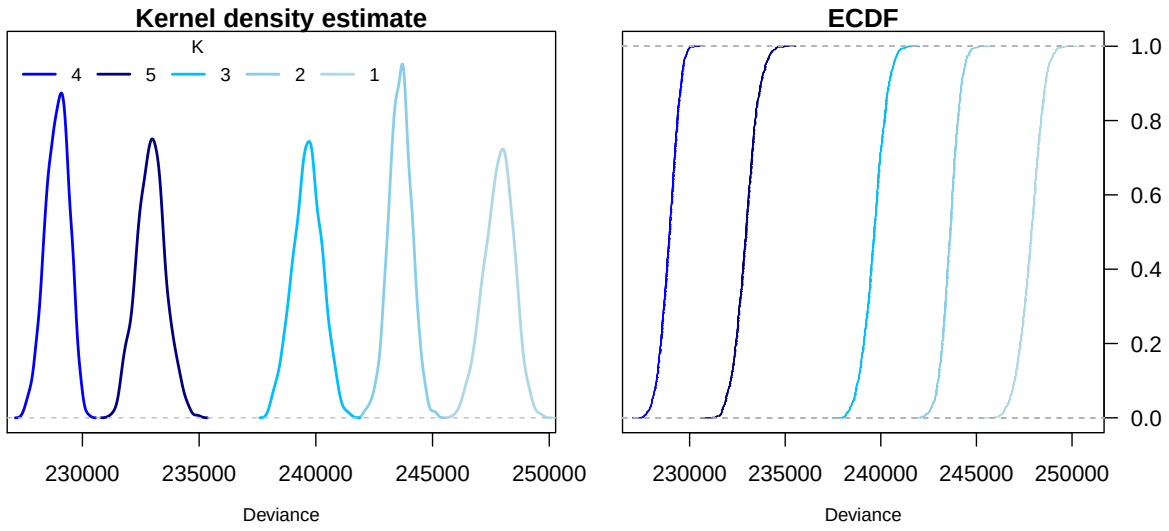


Fig. 5: Comparison of posterior distribution of deviances based on model with number of clusters $K = 1,2,3,4,5$.

in terms of classification (25 % of households remained unclassified) as the red, green and turquoise clusters do not substantially differ. To interpret such clusters more precisely we should not forget the other used outcomes. The resulting spline shapes for the affordability of week holiday and the financial burden of total housing cost can be found in the supplement, (Figures S12–S13).

Now we discuss the case $K = 5$ in more detail. As our goal is to find different patterns in evolution we may actually be even interested in blue and golden cluster of extreme antagonistic behaviour. These clusters 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 followed only for 4 consecutive years. Therefore, the blue cluster should be interpreted rather 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. Which leaves us with $K = 4$ solution being the most suitable for the overall interpretation.

Households in the violet cluster with exceptionally high income can also always afford to pay for one week of holiday abroad and do not find the housing cost to be a really heavy burden. On the other hand, turquoise cluster represents households of completely reverse characteristics - very low disposable income, inability to pay for a week holiday abroad and almost all of them struggle with payments for housing. The other two remaining clusters (red and green) share very similar and ordinary evolution of total disposable income, but can still be distinguished. See the proportions of categorical outcomes changing in time, especially the years 2010 and 2011 when the red cluster has the lowest percentage of households able to pay for week holiday, while the green cluster

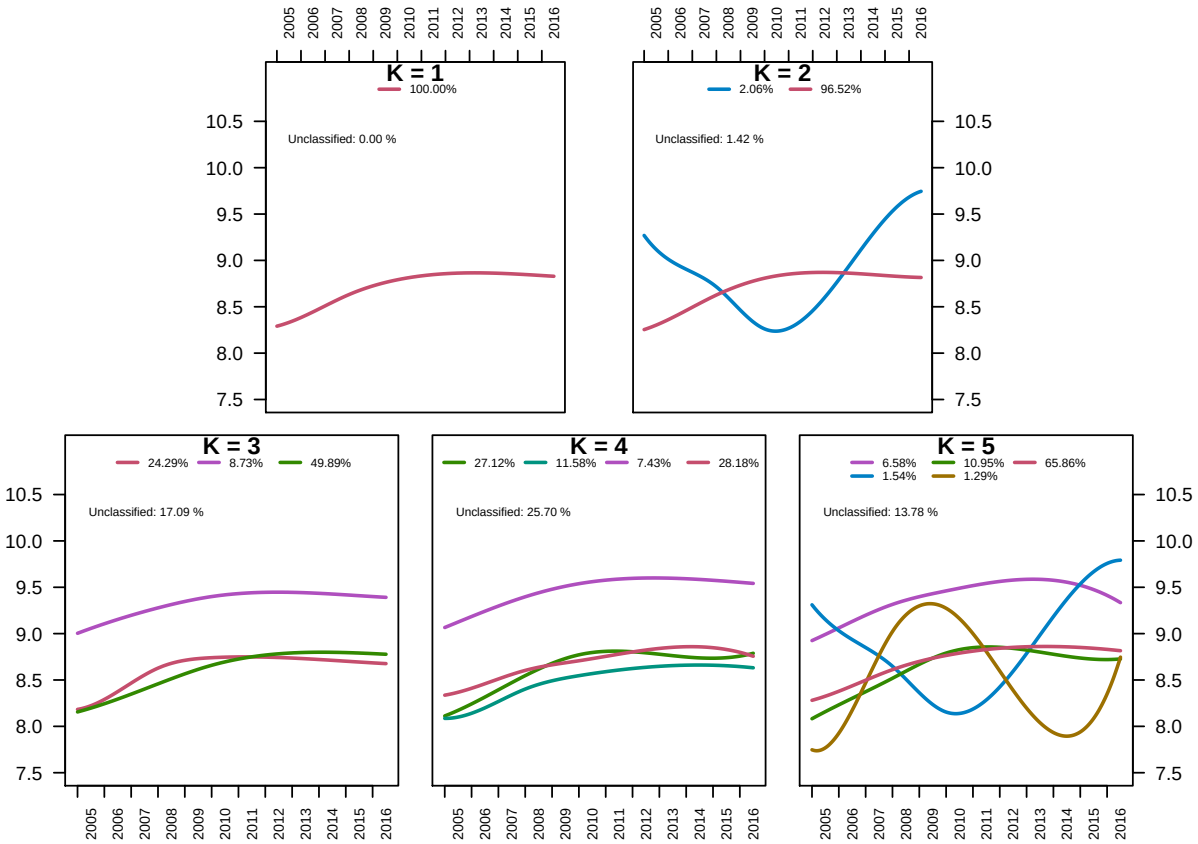


Fig. 6: Spline curves for logarithm of the total disposable household income of unit weighted family size for different choice of the number of clusters K.

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 that one big 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 allowed by our spline parametrization and the 4-year rotational panel invoked by the EU-SILC study.

7 Conclusion

In this paper we faced the problem of joint modelling of longitudinally measured numeric, ordinal and binary outcomes. We proposed to use multivariate linear mixed effects model on numeric and latent numeric outcomes corresponding to the categorical ones by exploiting the thresholding concept. Supposing all random effects to follow a joint normal distribution we enabled the outcomes to be related as it is common situation in real data analyses. On top of that, we enriched the model by construction of a mixture of such models allowing us to cluster individuals into several groups of various patterns in terms of time evolution or the variance covariance structure. By setting reasonable prior distributions for model and auxiliary parameters a hierarchical model was created and fully Bayesian approach was adopted and then executed using Gibbs sampling as one of the methods of MCMC. Appendix provides extensive derivations of full conditioned distributions used within the sampling mechanism. Sampled parameter values were not only used for the model estimation but for the calculation of classification probabilities as well, detailed derivation of which was provided in Section 4.7.

Proposed model and estimation method were tested in a simulation study with the aim to examine the ability to properly estimate model parameters and to correctly capture the patterns of each individual cluster. And indeed, we verified the consistency of parameter estimates even in the case of categorical variables modelled by simple thresholding. On the other hand, certain issues with the rate of convergence of variance matrix $\boldsymbol{\Sigma}$ or, especially, thresholds of latent numeric counterparts of ordinal variables have appeared. This motivates us to improve or even replace current Gibbs sampling with some more advanced MCMC techniques. That may include even automated selection of the total number of clusters K, where the inspiration for the future work comes from Neal (2000) and his use of Dirichlet process. Variance matrix $\boldsymbol{\Sigma}$ of random effects needs a careful attention as well as its dimension

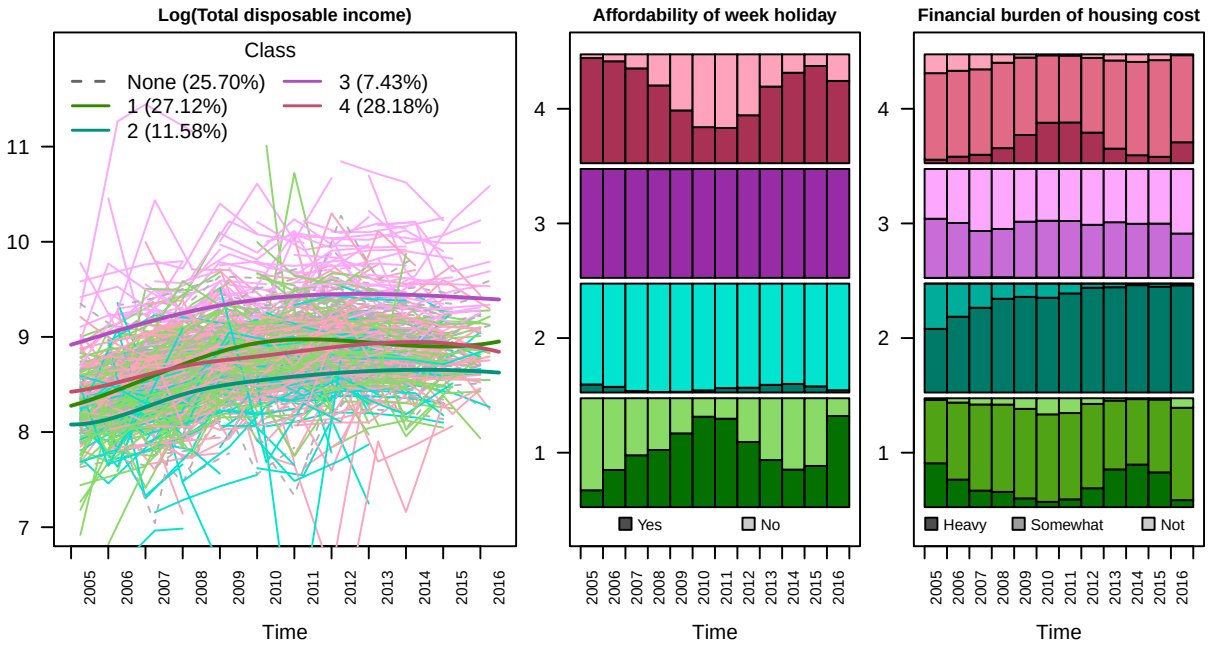


Fig. 7: Longitudinal profiles of numeric, binary and ordinal outcomes of $n = 1\,000$ randomly selected Czech households. Bold curves on the left represent the estimated conditional expectation of response within $K = 4$ discovered groups for a household of unit weighted family size. Categorical outcomes are presented by the proportions of categories in each year separately for the discovered groups. Some households remain unclassified.

rises with the number of modelled outcomes and the complexity of random effects structure. It may prove useful to abandon the complete generality and replace it with some commonly used block structures of variance matrices.

So far we examined the properties under fixed and minimal number of outcomes. Next work should be focused on the case of much higher number of measured outcomes and possibly on 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 education of the head of a household, ...), influence of which 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 week holiday abroad and self-evaluation of difficulty to pay for housing. Minorities of extremely high (7.43 %) or low (11.58 %) life standard are easily distinguishable unlike the other mid-class households of similar income and categorical proportions but different periods of increasing and decreasing tendencies. Using more than 4-cluster solution separates out households undergoing a huge progress or recession during a certain period of time. Though these findings may be of some interest, the corresponding patterns as a whole are unrealistic due to rotational design of the study, which demands only observations from four consecutive years, and hence, these clusters are irrelevant from the realistic point of view.

The whole estimation process was implemented completely from scratch using  for data preparation and processing while  functions were called during the sampling phase in hope to reduce the computation time. This combination truly proved to be more efficient then pure  implementation and resulted in a matter of minutes. The computation time, however, mainly depends on the number of subjects n - the larger it is, the higher number of latent parameters is to be sampled. For example, each of the cluster indicators U_i requires computation of K full conditional probabilities (38), which is still easily manageable as the latent numeric outcomes are at our disposal. This however, does not hold true for the classification probabilities (14) computed right after the sampling using an additional function within  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 version of `pmvnorm` function from `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. Therefore, some thinning methods were applied in order to obtain results in a still reasonable time.

Acknowledgement

This research was supported by the Czech Science Foundation (GAČR) grant 19-00015S.

Appendix

A Full-conditioned distributions in Gibbs sampling

In this section we derive full-conditioned distributions for all parameters $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ one by one. All derivations are based on viewing (26) as a function of parameter $\boldsymbol{\psi}$. The function on the right hand side of (26) can be decomposed into the following products:

$$
\begin{aligned}
p\left(\boldsymbol{\psi}\middle|\mathbb{Y},\mathscr{C};\boldsymbol{\Psi}_{-\boldsymbol{\psi}};\boldsymbol{\zeta}\right) \propto \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{OB}}\middle|\mathbb{Y}_i^{\star,\mathsf{OB}};\boldsymbol{\gamma}\right)\cdot \prod_{i=1}^{n} p\left(\mathbb{Y}_i^{\mathsf{N}},\mathbb{Y}_i^{\star,\mathsf{OB}}\middle|\mathscr{C}_i,\boldsymbol{b}_i;\boldsymbol{\beta}^{(U_i)},\boldsymbol{\tau}^{(U_i)}\right)\cdot \\
\cdot \prod_{i=1}^{n} p\left(\boldsymbol{b}_i\middle|\boldsymbol{\mu}^{(U_i)},\boldsymbol{\Sigma}^{-(U_i)}\right)\cdot \prod_{i=1}^{n} p\left(U_i|\boldsymbol{w}\right)\cdot \\
\cdot p\left(\boldsymbol{w}|\boldsymbol{\alpha}\right)\cdot p\left(\boldsymbol{\gamma}\middle|\gamma_1^{r^{\mathsf{O}}},r^{\mathsf{O}}\in\mathscr{R}^{\mathsf{O}}\right)\cdot p\left(\boldsymbol{\beta}|\boldsymbol{\tau};\boldsymbol{\beta}_0,\mathbb{D}\right)\cdot p\left(\boldsymbol{\tau}|a_1,a_2\right)\cdot \\
\cdot p\left(\boldsymbol{\mu}|\boldsymbol{\tau}_{\mathsf{R}};\boldsymbol{\mu}_0\right)\cdot p\left(\boldsymbol{\tau}_{\mathsf{R}}|a_3,a_4\right)\cdot p\left(\boldsymbol{\Sigma}^{-1}\middle|\mathbb{Q}^{-1};\nu_0\right)\cdot p\left(\mathbb{Q}^{-1}\middle|\mathbb{D}^{\mathbb{Q}},\nu_1\right). \qquad (36)
\end{aligned}
$$

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 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 be group-specific as well.

A.1 Probabilities $\boldsymbol{w}$

Parameter $\boldsymbol{w}$ of prior probabilities of being classified into certain categories, i.e. $w_k = \mathsf{P}(U_i = k)$, appear only in $p(U_i|\boldsymbol{w})$ and its prior distribution $p(\boldsymbol{w}|\boldsymbol{\alpha})$, therefore:

$$p(\boldsymbol{w}|\mathbb{Y},\mathscr{C};\boldsymbol{\Psi}_{-\boldsymbol{w}};\boldsymbol{\zeta}) \propto \prod_{i=1}^{n} p(U_i|\boldsymbol{w}) \cdot p(\boldsymbol{w}|\boldsymbol{\alpha})$$

$$p(\boldsymbol{w}|\boldsymbol{U};\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} w_k^{\sum_{i=1}^{n} \mathbb{1}(U_i=k)} \cdot \prod_{k=1}^{K} w_k^{\alpha_k - 1} = \prod_{k=1}^{K} w_k^{n^k(\boldsymbol{U})+\alpha_k-1},$$

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. Comparing this pdf with (19) we recognize the shape of pdf of Dirichlet distribution, thus,

$$\boldsymbol{w}|\mathbb{Y},\mathscr{C};\boldsymbol{\Psi}_{-\boldsymbol{w}};\boldsymbol{\zeta} \sim \mathsf{Dir}_K(\boldsymbol{n}(\boldsymbol{U}) + \boldsymbol{\alpha}), \tag{37}$$

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 (36) 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:

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

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

$$\mathsf{P}\left(U_i = k \middle| \mathbb{Y}_i^{\mathsf{N}}, \mathscr{C}_i; \mathbb{Y}_i^{\star,\mathsf{OB}}, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{w}\right) \propto w_k \cdot \prod_{r^{\mathsf{N}} \in \mathscr{R}^{\mathsf{Num}}} \left(\tau_{r^{\mathsf{N}}}^{(k)}\right)^{\frac{n_i}{2}} \cdot \left|\boldsymbol{\Sigma}^{-(k)}\right|^{\frac{1}{2}} \cdot$$

$$\cdot \exp\left\{-\frac{1}{2} \sum_{r^{\mathsf{N}} \in \mathscr{R}^{\mathsf{Num}}} \sum_{j=1}^{n_i} \tau_{r^{\mathsf{N}}}^{(k)} \left(y_{i,j}^{r^{\mathsf{N}}} - \eta_{i,j}^{(k),r^{\mathsf{N}}}\right)^2 - \frac{1}{2} \sum_{r^{\mathsf{OB}} \in \mathscr{R}^{\mathsf{OB}}} \sum_{j=1}^{n_i} \left(y_{i,j}^{\star,r^{\mathsf{OB}}} - \eta_{i,j}^{(k),r^{\mathsf{OB}}}\right)^2\right\} \cdot$$

$$\cdot \exp\left\{-\frac{1}{2}\left(\boldsymbol{b}_i - \boldsymbol{\mu}^{(k)}\right)^\top \boldsymbol{\Sigma}^{-(k)} \left(\boldsymbol{b}_i - \boldsymbol{\mu}^{(k)}\right)\right\}, \tag{38}$$

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 ith subject in 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 multivariate LME for both $\mathbb{Y}_i^{\mathsf{N}}$ and $\mathbb{Y}_i^{\star,\mathsf{OB}}$:

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

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

$$Y_{i,j}^{\star,r^{\mathsf{OB}}} \middle| Y_{i,j}^{r^{\mathsf{OB}}} = l, \boldsymbol{\gamma} \sim \mathsf{TN}\left(\eta_{i,j}^{(U_i),r^{\mathsf{OB}}}, 1, \gamma_l^{r^{\mathsf{OB}}}, \gamma_{l+1}^{r^{\mathsf{OB}}}\right). \tag{39}$$

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

Parameter $\boldsymbol{\gamma}$ affects (36) only in the thresholding phase and in prior distribution of $\boldsymbol{\gamma}$:

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

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

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

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

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

Thus, under the uniform prior (20) for $\boldsymbol{\gamma}^{r^{\mathsf{O}}}$ we get that individual thresholds $\gamma_l^{r^{\mathsf{O}}}$ are uniformly distributed on intervals given by maxima and minima of corresponding sets:

$$\gamma_l^{r^{\mathsf{O}}}\left|\boldsymbol{Y}^{r^{\mathsf{O}}};\boldsymbol{Y}^{\star,r^{\mathsf{O}}}\right.\sim\mathsf{Unif}\left[\max_{y\in\mathscr{Y}_{l-1}^{r^{\mathsf{O}}}}y,\ \min_{y\in\mathscr{Y}_l^{r^{\mathsf{O}}}}y\right],\qquad l=1,\ldots,L^{r^{\mathsf{O}}}.\tag{40}$$

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

Parameters $\boldsymbol{\tau}=\left\{\tau_{r^{\mathsf{N}}}^{(k)}:\ k=1,\ldots,K,\ r^{\mathsf{N}}\in\mathscr{R}^{\mathsf{Num}}\right\}$ are the inverse variance of errors of supposed LME models over numeric outcomes. The right-hand side of (36) includes $\boldsymbol{\tau}$ only in three factors - supposed LME for $\mathbb{Y}_i^{\mathsf{N}}$ and prior distribution of $(\boldsymbol{\beta},\boldsymbol{\tau})$:

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

From the structure of (7), (22) and (23) we see that individual $\tau_{r^{\mathsf{N}}}^{(k)}$ are distributed independently of each other (given all other information and parameters):

$$p\left(\tau_{r^{\mathsf{N}}}^{(k)}\middle|\boldsymbol{Y}^{r^{\mathsf{N}}},\mathscr{C}^{r^{\mathsf{N}}};\boldsymbol{U},\boldsymbol{b}^{r^{\mathsf{N}}};\boldsymbol{\beta}^{(k),r^{\mathsf{N}}};\boldsymbol{\beta}_0^{r^{\mathsf{N}}},\mathbb{D}^{r^{\mathsf{N}}},a_1,a_2\right)\propto\left(\tau_{r^{\mathsf{N}}}^{(k)}\right)^{\frac{1}{2}\sum_{i=1}^{n}n_i\mathbb{1}(U_i=k)+\frac{1}{2}d_{r^{\mathsf{N}}}^{\mathsf{F}}+a_1-1}\cdot$$
$$\cdot\exp\left\{-\tau_{r^{\mathsf{N}}}^{(k)}\left[\frac{1}{2}\sum_{i\in\mathscr{N}_k(\boldsymbol{U})}\sum_{j=1}^{n_i}\left(y_{i,j}^{r^{\mathsf{N}}}-\eta_{i,j}^{(k),r^{\mathsf{N}}}\right)^2+\frac{1}{2}\sum_{j=1}^{d_{r^{\mathsf{N}}}^{\mathsf{F}}}\frac{\left(\beta_j^{(k),r}-\beta_{0,j}^{r}\right)^2}{d_{j,j}^{r}}+a_2\right]\right\},\tag{41}$$

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 $\boldsymbol{Y}^{r^{\mathsf{N}}},\mathscr{C}^{r^{\mathsf{N}}}$ and current values of $\boldsymbol{U},\boldsymbol{b}^{r^{\mathsf{N}}}$ and $\boldsymbol{\beta}^{(k)}$ let us denote

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

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

Comparing (41) with (23) we see that

$$\tau_{r^{\mathsf{N}}}^{(k)}\left|\boldsymbol{Y}^{r^{\mathsf{N}}},\mathscr{C}^{r^{\mathsf{N}}};\boldsymbol{U},\boldsymbol{b}^{r^{\mathsf{N}}};\boldsymbol{\beta}^{(k),r^{\mathsf{N}}};\boldsymbol{\beta}_0^{r^{\mathsf{N}}},\mathbb{D}^{r^{\mathsf{N}}},a_1,a_2\right.\sim\Gamma\left(\widetilde{a}_1^{(k),r^{\mathsf{N}}},\widetilde{a}_2^{(k),r^{\mathsf{N}}}\right)\tag{42}$$

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

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

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

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

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

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

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

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

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

$$p\left(\boldsymbol{\beta}^{(k),r}\,\middle|\,\boldsymbol{Y}^r,\mathscr{C}^r;\boldsymbol{U},\boldsymbol{b}^r;\tau_r^{(k)};\boldsymbol{\beta}_0^r,\mathbb{D}^r\right)\propto$$
$$\exp\left\{-\frac{\tau_r^{(k)}}{2}\left(\boldsymbol{\beta}^{(k),r}-\widetilde{\boldsymbol{\beta}}^{(k),r}\right)^\top\left[\left(\mathbb{X}^r_{\mathscr{N}_k(\boldsymbol{U})}\right)^\top\mathbb{X}^r_{\mathscr{N}_k(\boldsymbol{U})}+(\mathbb{D}^r)^{-1}\right]\left(\boldsymbol{\beta}^{(k),r}-\widetilde{\boldsymbol{\beta}}^{(k),r}\right)\right\},\quad(43)$$

where

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

which compared to pdf of multivariate normal distribution yields

$$\boldsymbol{\beta}^{(k),r}\,\middle|\,\boldsymbol{Y}^r,\mathscr{C}^r;\boldsymbol{U},\boldsymbol{b}^r;\tau_r^{(k)};\boldsymbol{\beta}_0^r,\mathbb{D}^r\;\sim\;\mathsf{N}_{d_r^{\mathsf{F}}}\left(\widetilde{\boldsymbol{\beta}}^{(k),r},\frac{\left[\left(\mathbb{X}^r_{\mathscr{N}_k(\boldsymbol{U})}\right)^\top\mathbb{X}^r_{\mathscr{N}_k(\boldsymbol{U})}+(\mathbb{D}^r)^{-1}\right]^{-1}}{\tau_r^{(k)}}\right).\quad(44)$$

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 4.6.1. The derivation of full-conditioned distribution of this parameter is solely based on combining probability distribution functions in (24). Therefore,

$$p\left(\boldsymbol{\tau}_{\mathsf{R}}\,|\boldsymbol{\mu};\boldsymbol{\mu}_0,a_3,a_4\right)\propto\prod_{k=1}^{K}\prod_{j=1}^{d^{\mathsf{R}}}\left(\tau_{\mathsf{R},j}^{(k)}\right)^{a_3+\frac{1}{2}-1}\exp\left\{-\tau_{\mathsf{R},j}^{(k)}\left[a_4+\frac{1}{2}\left(\mu_j^{(k)}-\mu_{0,j}^{(k)}\right)^2\right]\right\}\quad(45)$$

and

$$\tau_{\mathsf{R},j}^{(k)}\left|\mu_j^{(k)};\mu_{0,j}^{(k)},a_3,a_4\right.\;\sim\;\Gamma\left(a_3+\frac{1}{2},a_4+\frac{1}{2}\left(\mu_j^{(k)}-\mu_{0,j}^{(k)}\right)^2\right)\quad(46)$$

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 (36) is in the case of this parameter simplified into

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

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

$$\begin{aligned}
p\left(\boldsymbol{\mu}^{(k)} \,\middle|\, \boldsymbol{U}, \boldsymbol{b}; \boldsymbol{\Sigma}^{-(k)}; \boldsymbol{\tau}_{\mathsf{R}}^{(k)}, \boldsymbol{\mu}_0^{(k)}\right) &\propto \prod_{i \in \mathscr{N}_k(\boldsymbol{U})} p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{-(k)}\right) \cdot p\left(\boldsymbol{\mu}^{(k)} | \boldsymbol{\tau}_{\mathsf{R}}^{(k)}; \boldsymbol{\mu}_0^{(k)}\right) \\
&\propto \exp\left\{-\frac{1}{2} \sum_{i \in \mathscr{N}_k(\boldsymbol{U})} \left(\boldsymbol{b}_i - \boldsymbol{\mu}^{(k)}\right)^{\top} \boldsymbol{\Sigma}^{-(k)} \left(\boldsymbol{b}_i - \boldsymbol{\mu}^{(k)}\right) - \frac{1}{2} \left(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}_0^{(k)}\right)^{\top} \mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right) \left(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}_0^{(k)}\right)\right\} \qquad (47) \\
&\propto \exp\left\{-\frac{1}{2} \left(\boldsymbol{\mu}^{(k)} - \widetilde{\boldsymbol{\mu}}^{(k)}\right)^{\top} \left[n^k(\boldsymbol{U}) \boldsymbol{\Sigma}^{-(k)} + \mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right)\right] \left(\boldsymbol{\mu}^{(k)} - \widetilde{\boldsymbol{\mu}}^{(k)}\right)\right\},
\end{aligned}$$

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

$$\boldsymbol{\mu}^{(k)} \left| \boldsymbol{U}, \boldsymbol{b}; \boldsymbol{\Sigma}^{-(k)}; \boldsymbol{\tau}_{\mathsf{R}}^{(k)}, \boldsymbol{\mu}_0^{(k)} \right. \sim \mathsf{N}_{d^{\mathsf{R}}}\left(\widetilde{\boldsymbol{\mu}}^{(k)}, \left[n^k(\boldsymbol{U}) \boldsymbol{\Sigma}^{-(k)} + \mathsf{diag}\left(\boldsymbol{\tau}_{\mathsf{R}}^{(k)}\right)\right]^{-1}\right) \qquad (48)$$

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 (36) shrinks into

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

Combining the two probability density functions in (25) we get

$$p\left(\mathbb{Q}^{-(k)} \,\middle|\, \boldsymbol{\Sigma}^{-(k)}; \nu_0, \nu_1, \mathbb{D}^{\mathbb{Q}}\right) \propto \left|\mathbb{Q}^{-(k)}\right|^{\frac{\nu_0 + \nu_1 - d^{\mathsf{R}} - 1}{2}} \exp\left\{-\mathsf{Tr}\left[\left(\boldsymbol{\Sigma}^{-(k)} + \left(\mathbb{D}^{\mathbb{Q}}\right)^{-1}\right) \mathbb{Q}^{-(k)}\right]\right\}, \qquad (49)$$

which compared to (25) resembles pdf of Wishart distribution. Therefore,

$$\mathbb{Q}^{-(k)} \left| \boldsymbol{\Sigma}^{-(k)}; \nu_0, \nu_1, \mathbb{D}^{\mathbb{Q}} \right. \sim \mathsf{W}_{d^{\mathsf{R}}}\left(\left[\boldsymbol{\Sigma}^{-(k)} + \left(\mathbb{D}^{\mathbb{Q}}\right)^{-1}\right]^{-1}, \nu_0 + \nu_1\right) \qquad (50)$$

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

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 36 only in pdf for random effects and in prior distribution of $\boldsymbol{\Sigma}^{-1}$:

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

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

$$\begin{aligned}p\left(\boldsymbol{\Sigma}^{-(k)}\left|\boldsymbol{U},\boldsymbol{b};\boldsymbol{\mu}^{(k)},\mathbb{Q}^{-(k)};\nu_0\right.\right)&\propto\left|\boldsymbol{\Sigma}^{-(k)}\right|^{\frac{n^k(\boldsymbol{U})+\nu_0-d^{\mathsf{R}}-1}{2}}\cdot\\&\cdot\exp\left\{-\frac{1}{2}\sum_{i\in\mathscr{N}_k(\boldsymbol{U})}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)^{\top}\boldsymbol{\Sigma}^{-(k)}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)-\mathsf{Tr}\left[\mathbb{Q}^{-(k)}\boldsymbol{\Sigma}^{-(k)}\right]\right\}\\&\propto\left|\boldsymbol{\Sigma}^{-(k)}\right|^{\frac{n^k(\boldsymbol{U})+\nu_0-d^{\mathsf{R}}-1}{2}}\exp\left\{-\mathsf{Tr}\left[\left(\mathbb{Q}^{-(k)}+\frac{1}{2}\sum_{i\in\mathscr{N}_k(\boldsymbol{U})}\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)\left(\boldsymbol{b}_i-\boldsymbol{\mu}^{(k)}\right)^{\top}\right)\boldsymbol{\Sigma}^{-(k)}\right]\right\},\end{aligned}\tag{51}$$

which compared to (25) resembles pdf of Wishart distribution. Therefore, independently for all $k=1,\ldots,K$

$$\boldsymbol{\Sigma}^{-(k)}\left|\boldsymbol{U},\boldsymbol{b};\boldsymbol{\mu}^{(k)},\mathbb{Q}^{-(k)};\nu_0\right.\;\sim\;\mathsf{W}_{d^{\mathsf{R}}}\left(\widetilde{\mathbb{Q}}^{(k)},n^k(\boldsymbol{U})+\nu_0\right),\tag{52}$$

where

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

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

The key role of our model is played by random effects $\boldsymbol{b}_i$, $i=1,\ldots,n$ that create linear predictors $\boldsymbol{\eta}_i^{(k),r}$, $k=1,\ldots,K$ and $r\in\mathscr{R}$. Probability density function of corresponding full-conditioned distribution is based on only two parts of the right-hand side of (36):

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

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 probability distribution function is of the shape

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

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

$$\widetilde{\boldsymbol{y}}_i=\begin{pmatrix}\vdots\\\sqrt{\tau_{r^{\mathsf{N}}}^{(k)}}\widetilde{\boldsymbol{y}}_i^{r^{\mathsf{N}}}\\\vdots\\\widetilde{\boldsymbol{y}}_i^{\star,r^{\mathsf{OB}}}\\\vdots\end{pmatrix},\quad\begin{matrix}r^{\mathsf{N}}\in\mathscr{R}^{\mathsf{Num}},\\r^{\mathsf{OB}}\in\mathscr{R}^{\mathsf{OB}},\end{matrix}\quad\widetilde{\mathbb{Z}}_i=\begin{pmatrix}\ddots&&&\\&\sqrt{\tau_{r^{\mathsf{N}}}^{(k)}}\mathbb{Z}_i^{r^{\mathsf{N}}}&&\\&&\ddots&\\&&&\mathbb{Z}_i^{r^{\mathsf{OB}}}\\&&&&\ddots\end{pmatrix}$$

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

$$\exp\left\{-\frac{1}{2}\left(\widetilde{\boldsymbol{b}}_i-\boldsymbol{b}_i\right)^\top\left[\widetilde{\mathbb{Z}}_i^\top\widetilde{\mathbb{Z}}_i+\boldsymbol{\Sigma}^{-(k)}\right]\left(\widetilde{\boldsymbol{b}}_i-\boldsymbol{b}_i\right)\right\}, \tag{53}$$

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 subject belonging to group $k=1,\ldots,K$ is

$$\boldsymbol{b}_i\left|\mathbb{Y}_i^{\mathsf{N}},\mathscr{C}_i;\mathbb{Y}_i^{\star,\mathsf{OB}},U_i;\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1}\right.\sim\mathsf{N}_{d^{\mathsf{R}}}\left(\widetilde{\boldsymbol{b}}_i,\left[\widetilde{\mathbb{Z}}_i^\top\widetilde{\mathbb{Z}}_i+\boldsymbol{\Sigma}^{-(k)}\right]^{-1}\right). \tag{54}$$

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, URL https://doi.org/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, URL http://www.jstor.org/stable/2532201

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 (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

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6(6):721–741, DOI 10.1109/TPAMI.1984.4767596

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

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

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

Robert CP (2001) The Bayesian Choice, 2nd edn. Springer-Verlag New York

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