Clusterwise multivariate regression of mixed-type panel data

Jan Vávra (vavraj@karlin.mff.cuni.cz)
Charles University
Arnošt Komárek
Charles University
Bettina Grün
Vienna University of Economics and Business
Gertraud Malsiner-Walli
Vienna University of Economics and Business

Research Article

Keywords: Multivariate longitudinal data, Mixed type outcome, Generalised linear mixed model (GLMM), Model-based clustering, Classification, Sparse finite mixture, EU-SILC

Posted Date: July 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1882841/v1

Clusterwise multivariate regression of mixed-type panel data

Jan Vávra[1*], Arnošt Komárek[1], Bettina Grün[2] and Gertraud Malsiner-Walli[2]

[1]Department of Probability and Mathematical Statistics, Charles University, Sokolovská 49/83, Prague, 186 75, Czech Republic.
[2]Institute for Statistics and Mathematics, Vienna University of Economics and Business, Welthandelsplatz 1, Vienna, 1020, Austria.

*Corresponding author(s). E-mail(s): vavraj@karlin.mff.cuni.cz;
Contributing authors: komarek@karlin.mff.cuni.cz;
bettina.gruen@wu.ac.at; gertraud.malsiner-walli@wu.ac.at;

Abstract

Multivariate panel data of mixed type are routinely collected in many different areas of application, often jointly with additional covariates which complicate the statistical analysis. Moreover, it is often of interest to identify unknown groups of units in a study population using such data structure, i.e., to perform clustering. In the Bayesian framework, we propose a finite mixture of multivariate generalised linear mixed effects regression models to cluster numeric, binary, ordinal and categorical panel outcomes jointly. The specification of suitable priors on the model parameters allows for convenient posterior inference based on Markov chain Monte Carlo (MCMC) sampling with data augmentation. The Bayesian approach allows to obtain both a classification of the subjects in the data and new subjects as well as cluster-specific parameter estimates. Finally, model estimation and selection of the number of data clusters are simultaneously performed when approximating the posterior for a single model using MCMC sampling without resorting to multiple model estimations. The performance of the proposed methodology is evaluated in a simulation study. Its application is illustrated on two data sets, one from a longitudinal patient study to infer prognosis groups, and a second one from the

Czech part of the EU-SILC survey where households are annually interviewed to obtain insights into changes in their financial capability.

Keywords: Multivariate longitudinal data, Mixed type outcome, Generalised linear mixed model (GLMM), Model-based clustering, Classification, Sparse finite mixture, EU-SILC

1 Introduction

Multivariate panel data containing several variables of different scale types are nowadays routinely collected in many different areas of application. However, panel data require specific statistical models and estimation methods in order to be able to harvest the full potential of this data collection mode (see, e.g., Fitzmaurice et al, 2008). Modelling panel data can be challenging for several reasons.

To begin with, the joint modelling of variables of different scale types usually assumes that the binary, general categorical and ordinal variables are manifestations of latent numeric variables (Agresti, 2013). This approach drastically increases the number of model parameters, especially for general categorical variables. Further, in multivariate data the variables included are in general not only of different scale types, but also have different roles attributed in the analysis. Some variables are considered to be the outcome or dependent variables whereas others are used as covariates to indicate how the conditional mean of the outcome variables varies in dependence of the covariate values. This naturally leads to a multivariate regression setup where the covariates as well as the outcome variables in the regression may be of different scale types, in particular, numeric, binary, ordinal and general categorical.

In panel data, several measurements are available for the same subjects and capturing the dependence between these measurements is of crucial importance. Repeated measurements for the same subjects are often imbalanced and induce within-subject correlations, only allowing for the identically and independently distribution assumption to hold on subject level. In a regression setting, random-effects models as proposed by Laird and Ware (1982) account for repeated measurements and within-subject correlation by introducing subject-specific coefficients capturing the between-subject variation through a normal distribution. These subject-specific coefficients are referred to as random effects, whereas the regression coefficients which are identical across subjects are referred to as fixed effects.

In a multivariate regression setting for panel data, random effects are specified for the regression model of each outcome variable. These random effects capture within-subject dependency and are not only correlated for each single regression model for each outcome variable, but are also correlated across the different regression models. This means that the random effects distribution needs to be specified to cover all random effects involved across the different

outcome variables, thus also increasing the dimension of the random effects (Fieuws and Verbeke, 2004).

Specifying only random effects following a normal distribution may be insufficient to capture and also characterise the between-subjects variability. Assuming that in fact groups of subjects exist where the effects on the outcome variables distinctively differ, leads to a model-based clustering problem (Fraley and Raftery, 2002). A finite mixture model of multivariate generalised linear mixed-effects regression models (GLMMs) allows to embed the clusterwise regression problem in a statistical modelling framework (Wedel and DeSarbo, 1995).

This proposed model for numeric, binary, ordinal and general categorical outcomes generalises and extends the models proposed in Fieuws and Verbeke (2004, 2006); Komárek and Komárková (2013); Komárek and Komárková (2014); Vávra and Komárek (2022). Fieuws and Verbeke (2004, 2006) considered multivariate mixed-effects regression models but did not account for unobserved heterogeneity using a mixture model; Komárek and Komárková (2013); Komárek and Komárková (2014) only allowed for binary and discrete count longitudinal outcomes in a model-clustering framework and Vávra and Komárek (2022) considered finite mixtures of multivariate classical normal mixed-effects regression models allowing only for numeric and both binary and ordinal outcomes (assuming the thresholding concept), excluding general categorical outcomes.

This paper proposes an approach which allows to discover clusters among multivariate longitudinal data of possibly different types by building on and combining several widely used methodologies. The model specification allows to combine an arbitrary number of numeric, binary, ordinal or general categorical outcome variables and to model them jointly by GLMMs. In Section 2, we first outline the multivariate GLMM approach which allows the joint modelling of mixed-type (numeric, binary, ordinal and general categorical) longitudinal data. In Section 3, we extend this model to allow for unobserved discrete heterogeneity using a mixture approach in the spirit of model-based clustering. This allows to classify and characterise subjects in particular due to distinctively different effects identified between the covariates and the outcome variables. Section 4 embeds the model within a Bayesian framework and outlines suitable prior specifications. In particular, a sparse finite mixture approach (Malsiner-Walli et al, 2016) allows to conveniently estimate the number of data clusters or groups. Section 5 provides the details of the Markov chain Monte Carlo (MCMC) algorithm for model estimation as well as the necessary post-processing steps to obtain an identified model. The simulation study in Section 6 evaluates the ability of the proposed model and inference methods to identify the true number of data clusters, to induce good cluster solutions and to characterise the clusters through the coefficient estimates. Sections 7 and 8 contain the analyses of two different data sets using the proposed approach: data from a medical study, where patients are monitored over

an extended time period with multiple laboratory measurements being available for each visit, and data from the EU-SILC (European Union Statistics on Income and Living Conditions) survey conducted in the Czech Republic from 2005 to 2018, where households are monitored for four consecutive years and information on their financial capabilities is collected. For both applications several variables naturally serve as outcome variables in a regression setting and identifying groups of patients or households with similar regression patterns is of core interest. Finally, Section 9 concludes.

2 Multivariate regression for mixed-type panel data

We assume that the multivariate panel data are composed of n subjects with n_i observations being available for each subject i. In addition R variables of mixed-type are considered as outcome variables in the regression models. These outcomes may have the following scale types: numeric, binary, ordinal or general categorical.

We define different index sets for the outcomes depending on the scale level such that $\mathcal{R}^{\mathsf{Num}}$ contains the indices of the numeric outcomes, $\mathcal{R}^{\mathsf{Bin}}$ those of the binary outcomes, $\mathcal{R}^{\mathsf{Ord}}$ those of the ordinal outcomes and $\mathcal{R}^{\mathsf{Cat}}$ those of the general categorical outcomes. This implies that $\mathcal{R} = \{1, \ldots, R\} = \mathcal{R}^{\mathsf{Num}} \cup \mathcal{R}^{\mathsf{Bin}} \cup \mathcal{R}^{\mathsf{Ord}} \cup \mathcal{R}^{\mathsf{Cat}}$.

In addition to the outcome observations $Y_{i,j}^r$ ($r = 1, \ldots, R$, $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$), for each subject i, additional observations are available which are used as covariates in the regression. We denote these additional variables, which are used as covariates in the regression for outcome variable r of subject i and its j-th observation, by $\boldsymbol{v}_{i,j}^r$. For subject i, let $\boldsymbol{Y}_i^r = \left(Y_{i,1}^r, \ldots, Y_{i,n_i}^r\right)^\top$ be the complete vector of all values of outcome r and $\mathcal{C}_i^r = \{\boldsymbol{v}_{i,1}^r, \ldots, \boldsymbol{v}_{i,n_i}^r\}$ the set of all covariates for outcome r. Combining them across the outcomes gives

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

which denotes all information (outcomes and covariate values) available for subject i. $\boldsymbol{Y}^r$ and $\mathcal{C}^r$ represent the complete information (outcome and covariate values) regarding one chosen outcome $r \in \mathcal{R}$ from all subjects. Finally, $\mathbb{Y}$ and $\mathcal{C}$ stand for all gathered information (all outcomes and covariate values) from all subjects.

The joint model for data (1) is built in the following way: For each outcome (each $r \in \mathcal{R}$) a generalised linear mixed model (GLMM) is assumed. A classical linear mixed model (LMM) is assumed for numeric outcomes and a logistic regression model with random effects is used for binary outcomes. The logistic regression model is extended for the ordinal and general categorical outcomes with more than two levels.

These individual regression models are combined into a multivariate model by assuming that the outcome variables given the regression models are independent between subjects and also within subjects conditional on the random effects of the mixed-effects models. However, the random effects are allowed to be correlated within subjects within and across the different outcomes. In this way correlation is induced between the outcomes given the regression models for each subject.

2.1 Generalised linear mixed models

A generalised linear mixed model is assumed for each outcome. This means that for each outcome a distribution from the exponential family is assumed as well as a link function which maps the linear predictor, determined by a linear combination of the fixed and random effects with their covariates, to the conditional mean of the outcome, given the covariate values. Thus, the linear predictor $\eta_{i,j}^r$ for observation j belonging to subject i specific to outcome r is given by the sum of

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

The linear predictor is thus given by $\eta_{i,j}^r = \eta_{i,j}^{\mathsf{F},r} + \eta_{i,j}^{\mathsf{R},r} = \left(\boldsymbol{x}_{i,j}^r\right)^\top \boldsymbol{\beta}_r + \left(\boldsymbol{z}_{i,j}^r\right)^\top \boldsymbol{b}_i^r$. The fixed-effects part $\eta_{i,j}^{\mathsf{F},r}$ captures the overall trend, and the random-effects part $\eta_{i,j}^{\mathsf{R},r}$ captures differences between subjects. While observations between different subjects are considered independent, the individual observations $j = 1, \ldots, n_i$ of subject i are assumed to be independent only given the random effects $\boldsymbol{b}_i^r$. Note that in the following notation, we may drop the three indices (i, j, r) at places where the notation would otherwise be unnecessarily complicated.

For each numeric outcome $r \in \mathcal{R}^{\mathsf{Num}}$, we assume the classical LMM (see Laird and Ware, 1982):

$$Y_{i,j}^r \;\middle|\; \boldsymbol{b}_i^r;\ \mathcal{C}_{i,j}^r \;\sim\; \mathsf{N}\left(\eta_{i,j}^r,\, \tau_r^{-1}\right),$$

where $\tau_r > 0$ is the precision (inverse variance) of the noise or regression model errors. The contribution of one numeric observation Y to the log-likelihood is given by:

$$\ell^{\mathsf{N}}\left(Y|\eta,\tau\right) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\tau - \frac{\tau}{2}\left(Y-\eta\right)^2.$$

Binary outcomes are assumed to follow a logistic regression model. The success probability is linked to the linear predictor using the inverse logit

function:

$$\mathsf{P}\left[Y = 1|\eta\right] = \mathsf{logit}^{-1}(\eta) = \frac{\exp\{\eta\}}{1 + \exp\{\eta\}}.$$

The contribution of one binary observation Y to the log-likelihood is then:

$$\ell^{\mathsf{B}}\left(Y|\eta\right) = Y\eta - \log\left(1 + \exp\{\eta\}\right).$$

Logit models for ordinal outcomes with $K > 2$ levels are obtained by parameterising the *cumulative* probabilities and linking them to the linear predictors using the inverse logit function (e.g., Hartzel et al, 2001, Section 2.2):

$$p_k := \mathsf{P}\left[Y > k|\eta, \boldsymbol{c}\right] = \mathsf{logit}^{-1}(\eta - c_k) \qquad \text{for any } k = 1, \ldots, K,$$

where $-\infty = c_0 < c_1 < \cdots < c_K = \infty$ are ordered intercepts that shift the linear predictor η and $\boldsymbol{c} = (c_k)_{k=0,\ldots,K}$. Note that for identifiability purposes the intercept term must *not* be included among the fixed effects in the logit models for ordinal outcomes. Also note that this model formulation is based on the proportional odds assumption; the log-odds differ only in the intercepts: $\log\left(\mathsf{P}\left[Y > k|\eta, \boldsymbol{c}\right] / \mathsf{P}\left[Y \leq k|\eta, \boldsymbol{c}\right]\right) = \eta - c_k, k = 1, \ldots, K$. For $K = 2$ this formulation is equivalent to the logistic regression model since a single free threshold c_1 is included in the model. This single threshold corresponds to the negative intercept term in logistic regression as $q_1 = 1 - p_1$ and $q_2 = p_1$. Using the notation $p_0 = \mathsf{P}\left[Y > 0\right] = 1$ and $p_K = \mathsf{P}\left[Y > K\right] = 0$, the probability of observing a value k is obtained as the difference between two consecutive cumulative probabilities:

$$q_k := \mathsf{P}\left[Y = k|\eta, \boldsymbol{c}\right] = \mathsf{P}\left[Y > k - 1|\eta, \boldsymbol{c}\right] - \mathsf{P}\left[Y > k|\eta, \boldsymbol{c}\right] = p_{k-1} - p_k.$$

The contribution of one ordinal observation Y to the log-likelihood is given by:

$$\ell^{\mathsf{O}}\left(Y = k|\eta, \boldsymbol{c}\right) = \log\left(q_k\right) = \log\left(p_{k-1} - p_k\right).$$

The logit parameterisation (Hartzel et al, 2001, Section 2.3) for a general categorical outcome with $K > 2$ levels requires a specific linear predictor η for each level $k = 1, \ldots, K$. Hence, the linear predictor for this outcome is the vector $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_K\}$ where each η_k is a linear combination of the same regressors with a different set of fixed effects $\boldsymbol{\beta}_{r,k}$ and random effects $\boldsymbol{b}^r_{i,k}$. The probability for level k is then obtained as the k-th element of the vector of probabilities obtained from transforming the linear predictor vector with the multivariate $\mathsf{softmax}$ function:

$$\mathsf{P}\left[Y = k|\boldsymbol{\eta}\right] = \mathsf{softmax}_k(\boldsymbol{\eta}) = \frac{\exp\{\eta_k\}}{\sum_{k'=1}^{K} \exp\{\eta_{k'}\}}.$$

This yields the probability ratios $\mathsf{P}[Y = k_1|\boldsymbol{\eta}] \,/\, \mathsf{P}[Y = k_2|\boldsymbol{\eta}] = \exp\{\eta_{k_1} - \eta_{k_2}\}$. For identifiability, the predictors $\boldsymbol{\eta}$ have to be restricted. We fix the last one to $\eta_K = 0$ by setting $\boldsymbol{\beta}_{r,K} = \mathbf{0}$ and $\boldsymbol{b}^r_{i,K} = \mathbf{0}$. This means it is sufficient to consider for $\boldsymbol{\eta}$ the $(K-1)$-dimensional vector containing $\{\eta_1, \ldots, \eta_{K-1}\}$. Imposing this restriction implies that the estimated regression coefficients capture the probability ratio between the k-th and the last category K. Hence, level K has a specific role and in general should correspond to some baseline level in applications. Note that under $K = 2$ this formulation reduces to the logistic regression assumed for binary outcomes. Then one has one actual predictor $\eta = \eta_1$ and fixes $\eta_2 = 0$. The contribution of one general categorical observation Y to the log-likelihood is given by:

$$\ell^{\mathsf{C}}(Y = k|\boldsymbol{\eta}) = \eta_k - \log\left(1 + \sum_{k'=1}^{K-1} \exp\{\eta_{k'}\}\right).$$

2.2 Combining the multivariate responses

In the GLMM framework, the random effects $\boldsymbol{b}^r_i$ capture the correlation between the outcome values observed for each subject i and outcome $r \in \mathcal{R}$ conditional on the regression model. In the multivariate setting with several different outcome variables, the random effects are also used to capture correlations between different outcome variables for a subject i. To this end, we suppose a joint multivariate distribution for all random effects similar to Fieuws and Verbeke (2004, 2006); Komárek and Komárková (2013); Komárek and Komárková (2014); Vávra and Komárek (2022).

Let us denote the set of fixed effects by $\boldsymbol{\beta} = \{\boldsymbol{\beta}_r, r \in \mathcal{R}\}$ and the overall vector of random effects for subject i by $\boldsymbol{b}_i = \{\boldsymbol{b}^r_i, r \in \mathcal{R}\}$. In the following we divide the vector $\boldsymbol{b}_i$ into subvectors depending on the type of outcomes to emphasise the resulting block structure. In particular, $\boldsymbol{b}^{\mathsf{N}}_i = \{\boldsymbol{b}^r_i, r \in \mathcal{R}^{\mathsf{Num}}\}$, $\boldsymbol{b}^{\mathsf{B}}_i = \{\boldsymbol{b}^r_i, r \in \mathcal{R}^{\mathsf{Bin}}\}$, $\boldsymbol{b}^{\mathsf{O}}_i = \{\boldsymbol{b}^r_i, r \in \mathcal{R}^{\mathsf{Ord}}\}$ and $\boldsymbol{b}^{\mathsf{C}}_i = \{\boldsymbol{b}^r_i, r \in \mathcal{R}^{\mathsf{Cat}}\}$. We will also use notation $\mathsf{type}(r) \in \{\mathsf{N}, \mathsf{B}, \mathsf{O}, \mathsf{C}\}$ for the corresponding type of outcome $r \in \mathcal{R}$.

The overall random effects vector $\boldsymbol{b}_i$ is now assumed to follow a centred multivariate normal distribution with a *general* covariance matrix, i.e., it is assumed

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}^{\mathsf{N}}_i \\ \boldsymbol{b}^{\mathsf{B}}_i \\ \boldsymbol{b}^{\mathsf{O}}_i \\ \boldsymbol{b}^{\mathsf{C}}_i \end{pmatrix} \overset{\text{iid}}{\sim} \mathsf{N}_{d^{\mathsf{R}}}\left(\mathbf{0},\, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{\mathsf{NN}} & \boldsymbol{\Sigma}^{\mathsf{NB}} & \boldsymbol{\Sigma}^{\mathsf{NO}} & \boldsymbol{\Sigma}^{\mathsf{NC}} \\ \boldsymbol{\Sigma}^{\mathsf{BN}} & \boldsymbol{\Sigma}^{\mathsf{BB}} & \boldsymbol{\Sigma}^{\mathsf{BO}} & \boldsymbol{\Sigma}^{\mathsf{BC}} \\ \boldsymbol{\Sigma}^{\mathsf{ON}} & \boldsymbol{\Sigma}^{\mathsf{OB}} & \boldsymbol{\Sigma}^{\mathsf{OO}} & \boldsymbol{\Sigma}^{\mathsf{OC}} \\ \boldsymbol{\Sigma}^{\mathsf{CN}} & \boldsymbol{\Sigma}^{\mathsf{CB}} & \boldsymbol{\Sigma}^{\mathsf{CO}} & \boldsymbol{\Sigma}^{\mathsf{CC}} \end{pmatrix}\right),$$

where $d^{\mathsf{R}} = \sum_{r \in \mathcal{R}} d^{\mathsf{R}}_r$ is the total dimension of $\boldsymbol{b}_i$ and $\boldsymbol{\Sigma} > 0$ is the positive definite covariance matrix of the random effects. A general structure is assumed for this matrix thus allowing to capture arbitrary within-subject dependencies between the different outcomes.

Throughout the manuscript, the notation $p(\cdot\,|\,\cdot)$ and $\ell(\cdot\,|\,\cdot)$ indicate a conditional probability distribution function and its logarithm, respectively. The unknown parameters of the model consist of the fixed effects $\boldsymbol{\beta}$, the covariance matrix $\boldsymbol{\Sigma}$, the precisions of the error terms of the LMMs for numeric outcomes $\boldsymbol{\tau} = \left\{\tau_r, r \in \mathcal{R}^{\mathsf{Num}}\right\}$ and the ordered intercepts $\boldsymbol{c} = \left\{\boldsymbol{c}_r, r \in \mathcal{R}^{\mathsf{Ord}}\right\}$. The random effects $\boldsymbol{b}_i$ are unknown as well. However, these are considered latent variables that are integrated out to obtain the likelihood.

The multivariate GLMM implies that the i-th subject has the following likelihood contribution

$$p(\mathbb{Y}_i|\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\tau},\boldsymbol{c};\ \mathcal{C}_i) = \int \underbrace{\prod_{r=1}^{R}\prod_{j=1}^{n_i} \exp\left\{\ell^{\mathsf{type}(r)}\left(Y_{i,j}^r|\boldsymbol{\eta}_{i,j}^r,\boldsymbol{\tau},\boldsymbol{c}\right)\right\}}_{p(\mathbb{Y}_i|\,\boldsymbol{b}_i,\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{c};\ \mathcal{C}_i\,)} \cdot \underbrace{|2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{b}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{b}_i\right\}}_{p(\boldsymbol{b}_i|\,\boldsymbol{\Sigma}\,)} \mathrm{d}\boldsymbol{b}_i. \quad (2)$$

The integral (2) can be expressed in closed form only if all outcomes are numeric, i.e., $\mathcal{R} = \mathcal{R}^{\mathsf{Num}}$. Otherwise, numerical methods such as Adaptive Gaussian Quadrature (AGQ) have to be used to approximate the integral (for more details see Section 3.2).

3 Extending to model-based clustering

The multivariate regression for mixed-type panel data proposed so far accounts for subject-specific slight differences and has the fixed effects capturing an overall population effect. This specification assumes that all heterogeneity in the outcome variables can be essentially captured by the available covariates. However, in case of unobserved heterogeneity, i.e., if the population in fact contains several groups where different multivariate regression models apply with varying effects and conditional distributions, this model formulation is insufficient and extension to a mixture model warranted.

A mixture model enables a clusterwise regression setup where based on a model-based clustering approach subjects are classified into groups having similar regression effects. Such a mixture model allows to classify available subjects as well as new subjects into groups. A group-specific analysis is helpful for a better understanding how the effects of the covariates differ across groups in the population.

3.1 Creating a mixture distribution

Unobserved heterogeneity refers to the fact that there exists a discrete variable $U_i \in \left\{1, \ldots, G\right\}$ which represents the unobserved group-allocation indicator for subject i $(i = 1, \ldots, n)$. Within each group g, the model for subject i

is given by $p\big(\mathbb{Y}_i \,\big|\boldsymbol{\beta}^{(g)}, \boldsymbol{\Sigma}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{c}^{(g)};\ \mathcal{C}_i\big)$ of the form (2) with group-specific parameters being inserted as indicated by the superscript (g).

This formulation assumes that all parameters vary across groups. However, in general one splits the set of all unknown parameters into a set of parameters which are common to all groups, which we will denote by $\boldsymbol{\zeta}$ in the following, and a set of parameters which are group-specific, i.e., $\boldsymbol{\zeta}^{(g)}$ for the parameters specific to group g. The combination of all group-specific parameters is denoted by $\boldsymbol{\zeta}^{1:G} = \big\{\boldsymbol{\zeta}^{(g)}, g = 1, \ldots, G\big\}$.

This formulation implies that the assumed conditional probability distribution function of the ith subject's outcomes given the group allocation U_i is

$$p\left(\mathbb{Y}_i \,\middle|\, U_i = g,\, \boldsymbol{\zeta},\, \boldsymbol{\zeta}^{(g)};\ \mathcal{C}_i\right) \overset{(2)}{=} \int \underbrace{\prod_{r=1}^{R}\prod_{j=1}^{n_i} \exp\left\{\ell^{\mathsf{type}(r)}\left(Y_{i,j}^{r} | \boldsymbol{\eta}_{i,j}^{r},\, \boldsymbol{\zeta},\, \boldsymbol{\zeta}^{(g)}\right)\right\}}_{p\left(\mathbb{Y}_i \,\middle|\, \boldsymbol{b}_i,\, U_i=g,\, \boldsymbol{\zeta},\, \boldsymbol{\zeta}^{(g)};\ \mathcal{C}_i\right)} \cdot \underbrace{\left|2\pi\boldsymbol{\Sigma}^{(g)}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{b}_i^{\top}\boldsymbol{\Sigma}^{-(g)}\boldsymbol{b}_i\right\}}_{p\left(\boldsymbol{b}_i \,\middle|\, U_i=g,\, \boldsymbol{\Sigma}^{(g)}\right)} \mathrm{d}\boldsymbol{b}_i, \quad (3)$$

where we use the notation $\boldsymbol{\Sigma}^{-(g)}$ for the inverse matrix of $\boldsymbol{\Sigma}^{(g)}$, i.e., the precision matrix.

Let $w_g = \mathsf{P}\big(U_i = g \,\big|\, \boldsymbol{w}\big) \in (0, 1)$, $g = 1, \ldots, G$, $\sum_{g=1}^{G} w_g = 1$, be the (unknown) group sizes, with $\boldsymbol{w} := \big(w_1, \ldots, w_G\big)$. Integrating out the unobserved group membership U_i, the mixture distribution for the observed outcomes $\mathbb{Y}_i$ of a single subject i given covariates and model parameters $\boldsymbol{\theta} = \big\{\boldsymbol{w}, \boldsymbol{\zeta}, \boldsymbol{\zeta}^{1:G}\big\}$ corresponds to

$$p\left(\mathbb{Y}_i \middle|\ \boldsymbol{\theta};\ \mathcal{C}_i\right) = \sum_{g=1}^{G} w_g \int p\left(\mathbb{Y}_i \middle|\ \boldsymbol{b}_i, U_i = g, \boldsymbol{\theta};\ \mathcal{C}_i\right)\, p\left(\boldsymbol{b}_i \middle|\ U_i = g, \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{b}_i. \quad (4)$$

i.e., a mixture distribution which consists of G components with component weights $\boldsymbol{w}$ and component distributions resulting from the integration.

3.2 Classifying (new) observations

Given the model and its parameters $\boldsymbol{\theta}$, one can assign observed subjects to groups based on their classification probabilities, i.e., the a-posteriori probabilities to be from each of the group. In this way a partition of the available subjects is obtained which usually is of core interest in clustering applications.

The classification probability or the conditional probability of subject i to be from group g given the observed data is provided by the Bayes rule:

$$u_{i,g}(\boldsymbol{\theta}) := \mathsf{P}\left[U_i = g \mid \mathbb{Y}_i, \boldsymbol{\theta};\ \mathcal{C}_i\right] = \frac{w_g\, p\left(\mathbb{Y}_i \,\middle|\, U_i = g, \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(g)};\ \mathcal{C}_i\right)}{\sum_{g'=1}^{G} w_{g'}\, p\left(\mathbb{Y}_i \middle| U_i = g', \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(g')};\ \mathcal{C}_i\right)}. \quad (5)$$

Expression (5) can also be expressed as

$$u_{i,g}(\boldsymbol{\theta}) = \frac{w_g \left|\boldsymbol{\Sigma}^{(g)}\right|^{-\frac{1}{2}} \int \exp\left\{h\left(\boldsymbol{b}_i;\ \mathbb{Y}_i, \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(g)};\ \mathcal{C}_i\right)\right\} \mathrm{d}\boldsymbol{b}_i}{\sum_{g'=1}^{G} w_{g'} \left|\boldsymbol{\Sigma}^{(g')}\right|^{-\frac{1}{2}} \int \exp\left\{h\left(\boldsymbol{b}_i;\ \mathbb{Y}_i, \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(g')};\ \mathcal{C}_i\right)\right\} \mathrm{d}\boldsymbol{b}_i},$$

where

$$h(\boldsymbol{b}_i) := h\left(\boldsymbol{b}_i;\ \mathbb{Y}_i, \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(g)};\ \mathcal{C}_i\right) = \sum_{r\in\mathcal{R}} \sum_{j=1}^{n_i} \ell^{\mathsf{type}(r)}\left(Y_{i,j}^r | \boldsymbol{\eta}_{i,j}^r, \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(g)}\right) - \frac{1}{2}\boldsymbol{b}_i^\top \boldsymbol{\Sigma}^{-(g)} \boldsymbol{b}_i.$$

In the special case where $\mathcal{R} = \mathcal{R}^{\mathsf{Num}}$ (i.e., only normally distributed numeric outcomes are included in the model) the integrals $\int \exp\{h(\boldsymbol{b}_i)\}\,\mathrm{d}\boldsymbol{b}_i$ are available in closed form. In general, however, when also categorical outcomes are present, we have to approximate these integrals, which is achieved by the methodology summarised by Pinheiro and Chao (2006).

A rather crude approximation is possible using the Laplacian approximation, which relies on the Taylor expansion of the function h around its stationary point $\widehat{\boldsymbol{b}}_i^{(g)}$ that can be found by Newton–Raphson's method together with the negative Hessian matrix $H^{(g)}$ at this point.* The integral is approximated by $\int \exp\{h(\boldsymbol{b}_i)\}\,\mathrm{d}\boldsymbol{b}_i \approx \exp\{h(\widehat{\boldsymbol{b}}_i^{(g)})\} \cdot |H^{(g)}|^{-\frac{1}{2}}$ up to a multiplicative constant, yielding

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

More precise approximations can be obtained via Adaptive Gaussian Quadrature (AGD) which generalises the Laplacian approximation and evaluates function h at more than just one single point. The grid of d^{R}-dimensional points $\boldsymbol{z_j}$ consisting of the roots z_{j_l} of the Hermite polynomial of order N_{GQ} is scaled and translated into $\widetilde{\boldsymbol{b}}_{i,\boldsymbol{j}}^{(g)} = \widehat{\boldsymbol{b}}_i^{(g)} + \left(H^{(g)}\right)^{-\frac{1}{2}} \boldsymbol{z_j}$. These roots are tied

*Note that these derivatives of function h are also required for sampling $\boldsymbol{b}_i$ from its full-conditioned distribution, see Appendix B.3.

with weights $W_{\boldsymbol{j}} = \exp\{\|\boldsymbol{z}_{\boldsymbol{j}}\|^2\} \prod_{l=1}^{d^R} v_{j_l}$, where v_{j_l} corresponds to the weight of the root z_{j_l}. The classification probabilities are approximated by

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

In applications we recommend to use a rather low value of N_{GQ}, since there are $N_{GQ}^{d^R}$ summands to be evaluated, which can be costly to compute.

4 Bayesian modelling with suitable prior specifications

The model parameters $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{\zeta}, \boldsymbol{\zeta}^{1:G}\}$ imply the likelihood

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G} w_g p\left(\mathbb{Y}_i \,\middle|\, U_i = g,\, \boldsymbol{\zeta},\, \boldsymbol{\zeta}^{(g)};\; \mathcal{C}_i\right) \right\}. \tag{6}$$

In the following we will pursue a Bayesian approach and determine the posterior distribution of the model parameters. Compared to maximum likelihood estimation, the Bayesian framework allows to regularise the mixture likelihood through the prior specification, eases inference through data augmentation and enables convenient estimation of the number of data clusters.

The Bayesian framework and the related MCMC methodology allow for full exploitation of the hierarchical structure of the model. The integration with respect to the unobserved quantities $(U_i, \boldsymbol{b}_i)$ is elegantly avoided by data augmentation and the sampling mechanism. This also applies potentially to missing outcome values, for which a predictive distribution can be obtained simultaneously with the model estimation as long as all covariates are at disposal. Such an approach allows to retain more observations compared to a complete case analysis and thus is more informative. Moreover, the likelihood (6) is regularised by setting up convenient prior distributions over model parameters. Additionally, the Bayesian framework enables the estimation of the number of clusters in the data by specifying a suitable prior distribution on the component weights $\boldsymbol{w}$.

We employ the usual prior specification used in Bayesian model-based clustering where exchangeable priors are imposed on the components and their weights, and they are only potentially coupled through the use of hierarchical priors. Assuming that the priors on the component weights and the covariance matrices depend on hyperparameters e_0 and $\mathbb{Q}$, respectively, the joint prior

distribution imposed decomposes into

$$p(\boldsymbol{\theta},\, e_0,\, \mathbb{Q}) = p(\boldsymbol{w}|e_0)\, p(e_0)\, p\big(\boldsymbol{\beta} \,\big|\, \boldsymbol{\tau}\big)\, p(\boldsymbol{\tau})\, p(\boldsymbol{c})\, p(\boldsymbol{\Sigma}|\mathbb{Q})\, p(\mathbb{Q}).$$

In the following, the suitable prior specifications for this modelling approach in a model-based clustering context are discussed in detail.

4.1 Prior setting for the component distributions

In Bayesian mixture modelling, improper priors are not feasible for the component distributions because they induce an improper posterior as components have a positive probability of containing not a single observation (Roeder and Wasserman, 1997). Thus proper priors need to be selected. Also, selecting priors which impose a certain amount of regularisation eliminates spurious modes from the likelihood. In the regression case we thus standardise covariates prior to the analysis and impose coefficient priors gauged to the unit scale.

In the following, we only discuss the priors specified when all parameters characterising the component distributions are group-specific. Alternatively, one could also split these parameter vectors into a sub-vector containing the group-specific parameters and a sub-vector containing the parameters which are identical across groups. In this case, the priors have to be suitably modified, but the specifications are in an analogous way.

Priors on the fixed-effects coefficients and the precisions

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

The regression coefficients for the binary, ordinal and general categorical outcomes, i.e., $\beta_{r,j}^{(g)}, r \in \mathcal{R}^{\mathsf{Bin}} \cup \mathcal{R}^{\mathsf{Ord}}$ and $\beta_{r,k,j}^{(g)}, r \in \mathcal{R}^{\mathsf{Cat}}$ are also assumed to be a-priori independent and follow an analogous normal distribution $\mathsf{N}\left(\beta_{0,r,j},\, d_{j,j}^r\right)$, where, however, no precision parameter $\boldsymbol{\tau}$ is involved.

Regarding the ordered intercepts $\boldsymbol{c}_r^{(g)}$ estimated for ordinal outcomes, i.e., $r \in \mathcal{R}^{\mathsf{Ord}}$, the prior is not specified for them directly, but for transformed quantities. The $(K-1)$-dimensional ordered intercepts $\left(c_{r,1}^{(g)}, \ldots, c_{r,K-1}^{(g)}\right)$ are transformed into probabilities $\left(\pi_{r,1}^{(g)}, \ldots, \pi_{r,K}^{(g)}\right)$:

$$\begin{aligned} \pi_{r,k}^{(g)} &= \mathsf{P}\left[Y_{i,j}^r = k \,\middle|\, \boldsymbol{b}_i = \mathbf{0},\, U_i = g,\, \boldsymbol{x}_{i,j}^r = \mathbf{0}\right] \\ &= \mathsf{logit}^{-1}\left(c_{r,k}^{(g)}\right) - \mathsf{logit}^{-1}\left(c_{r,k-1}^{(g)}\right), \end{aligned} \tag{7}$$

$$c_{r,k}^{(g)} = \log\left(\frac{\pi_{r,1}^{(g)} + \cdots + \pi_{r,k}^{(g)}}{\pi_{r,k+1}^{(g)} + \cdots \pi_{r,K}^{(g)}}\right).$$

The prior distribution is then specified for the probabilities $\pi_{r,k}^{(g)}$ for all outcomes $r \in \mathcal{R}^{\mathsf{Ord}}$ using a product of Dirichlet distributions:

$$p(\boldsymbol{\pi}) \propto \prod_{r \in \mathcal{R}^{\mathsf{Ord}}} \prod_{g=1}^{G} \prod_{k=1}^{K} \left(\pi_{r,k}^{(g)}\right)^{\alpha_{r,k}-1}, \tag{8}$$

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

The precision parameters $\boldsymbol{\tau}_r^{(g)}$ for numeric outcomes are assumed to follow independent Gamma priors $\tau_r^{(g)} \sim \Gamma(\alpha_1, \alpha_2)$ with shape $\alpha_1 > 0$ and rate $\alpha_2 > 0$. For calculations in the later applications, we use $\alpha_1 = \alpha_2 = 1$.

Priors on the random effects parameters

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

4.2 Prior setting for the component weights: Sparse finite mixtures

Following the usual Bayesian mixture modelling specification, we impose a symmetric Dirichlet prior on the component weights $\boldsymbol{w}$:

$$\boldsymbol{w}|e_0 \sim \mathsf{Dir}_G\left((e_0, \ldots, e_0)\right) \equiv \mathsf{Dir}_G\left(e_0\right) \tag{9}$$

with probability density function

$$p(\boldsymbol{w}|e_0) = \frac{\Gamma(G \cdot e_0)}{(\Gamma(e_0))^G} \prod_{g=1}^{G} w_g^{e_0}.$$

The specification of e_0 is crucial depending on whether one assumes a-priori that subjects from all components are contained in the data set with a high

probability. We denote by G_+ the number of clusters from which subjects are generated in the given data set. The choice of e_0 controls the prior probability of $G_+ < G$. This probability is high when e_0 is small because the Dirichlet prior then puts a lot of mass on the boundary regions of the simplex and many of the G weights are small a-priori. For large values of e_0 one has with high probability $G = G_+$ a-priori, i.e., the number of data clusters coincides with the number of components specified.

We follow Frühwirth-Schnatter (2011) when specifying e_0 to take into account if the number of groups in the data set is known or should be estimated from the data. In case the number of groups in the data set are a-priori known, one would thus set G equal to this number and use a rather large value for e_0. By contrast, in case one needs to estimate G_+ from the data set, it is convenient to pursue the *sparse finite mixture* approach proposed by Malsiner-Walli et al (2016). This approach consists of selecting a large, fixed value for the number of components G such that G clearly exceeds the number of clusters in the data. In combination with a small value for e_0, one achieves that a-priori $G_+ \ll G$ and one may obtain a posterior distribution for G_+ which combines the prior specification of a small number of data clusters with the information on the cluster structure contained in the data.

To attenuate the influence of a specific choice of e_0, we assign a Gamma prior on e_0:

$$e_0 | a_e, b_e \sim \Gamma(a_e, b_e) \tag{10}$$

with probability density function

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

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

5 Bayesian inference

For Bayesian inference, we exploit the ideas of Bayesian data augmentation (Tanner and Wong, 1987) while considering all latent quantities, i.e., the component allocations $\boldsymbol{U} := \{U_i, i = 1, \ldots, n\}$, the random effect vectors $\boldsymbol{b} := \{\boldsymbol{b}_i, i = 1, \ldots, n\}$ and the missing outcome values denoted by $\mathbb{Y}^{\mathsf{mis}}$ as additional latent variables included in the posterior distribution. The model specified in Sections 2 and 3 results in the following joint distribution of the complete set of outcomes $\mathbb{Y} = (\mathbb{Y}^{\mathsf{obs}}, \mathbb{Y}^{\mathsf{mis}})$ divided into the observed data $\mathbb{Y}^{\mathsf{obs}}$ and the missing data $\mathbb{Y}^{\mathsf{mis}}$ together with the latent variables $\{\boldsymbol{U}, \boldsymbol{b}\}$, the model

parameters $\boldsymbol{\theta}$ and the hyperparameters e_0 and $\mathbb{Q}$:

$$
\begin{aligned}
p\big(\mathbb{Y},\ \boldsymbol{U},\ \boldsymbol{b},\ \boldsymbol{\theta}, e_0, \mathbb{Q};\ \mathcal{C}\big) \ &= \qquad (11)\\
&= \left[\prod_{i=1}^{n} p\big(\mathbb{Y}_i \mid \boldsymbol{b}_i,\ U_i,\ \boldsymbol{\theta};\ \mathcal{C}_i\big)\, p\big(\boldsymbol{b}_i \,\big|\, U_i, \boldsymbol{\theta}\big)\, p\big(U_i \,\big|\, e_0\big)\right] p(\boldsymbol{\theta}|\mathbb{Q}) p(\mathbb{Q}) p(e_0)\\
&= \left[\prod_{i=1}^{n} p\big(\mathbb{Y}_i \mid \boldsymbol{b}_i,\ \boldsymbol{\beta}^{(U_i)},\ \boldsymbol{\tau}^{(U_i)},\ \boldsymbol{c}^{(U_i)};\ \mathcal{C}_i\big)\, p\big(\boldsymbol{b}_i \,\big|\, \boldsymbol{\Sigma}^{(U_i)}\big)\, w_{U_i}\right]\\
&\qquad\qquad p(\boldsymbol{\theta}|\mathbb{Q}) p(\mathbb{Q}) p(e_0),
\end{aligned}
$$

where $p(\boldsymbol{\theta}|\mathbb{Q})$ is the prior distribution of the model parameters given the scale matrix $\mathbb{Q}$, $p(\mathbb{Q})$ is the prior for the scale matrix and $p(e_0)$ is the prior of the Dirichlet parameter e_0.

5.1 MCMC algorithm

The posterior distribution $p\big(\boldsymbol{\theta},\ \boldsymbol{U},\ \boldsymbol{b},\ e_0,\ \mathbb{Q},\ \mathbb{Y}^{\mathsf{mis}} \big|\ \mathbb{Y}^{\mathsf{obs}};\ \mathcal{C}\big)$ is estimated using MCMC sampling (Brooks et al, 2011). In particular, we adopt the classical Gibbs sampling scheme wherever possible. Due to the (semi)-conjugate choices of prior distributions, the full-conditioned distributions of $\boldsymbol{\beta}_r^{(g)}, r \in \mathcal{R}^{\mathsf{Num}}$, $\boldsymbol{\tau}^{(g)}$, $\boldsymbol{\Sigma}^{(g)}, \mathbb{Q}$, $\boldsymbol{w}$, $\boldsymbol{U}$ and $\mathbb{Y}^{\mathsf{mis}}$ belong to well known distributional families, for which efficient and straightforward sampling mechanisms are available, requiring only updates of the parameters. This is not the case for $\boldsymbol{\beta}_r^{(g)}, r \in \mathcal{R} \setminus \mathcal{R}^{\mathsf{Num}}$, $\boldsymbol{c}$, $\boldsymbol{b}$ and e_0, which are sampled using a Metropolis proposal step. More details can be found in the Appendix B.

The sampling algorithm can be summarised as follows:

1) Choose an initial partition $\mathcal{P}$, values for the unknown parameters and repeat the Steps 2) – 7).
2) Sample the missing outcome values $\mathbb{Y}^{\mathsf{mis}}$ according to the data-generating process implied by the specified model.
3) Sample the component-specific parameters $\boldsymbol{\zeta}^{(g)}$ for $g = 1, \ldots, G$:
   a) If $n^{(g)} > 0$ (non-empty component): sample the parameters from full-conditioned distributions (directly or using a Metropolis step) using the observations of the subjects currently assigned to cluster g.
   b) If $n^{(g)} = 0$ (empty component): sample the parameters from their prior distributions (directly or using a Metropolis step).

   $n^{(g)}$ denotes the number of subjects assigned to component g.
4) Sample the parameters $\boldsymbol{\zeta}$ which are identical across components and the scale matrix $\mathbb{Q}$ from their full-conditioned distributions (directly or using a Metropolis step).
5) Sample the component weights $\boldsymbol{w}$ from the Dirichlet distribution given by $\mathsf{Dir}_G\,(\boldsymbol{n} + e_0\mathbf{1})$, where $\boldsymbol{n} = (n^{(1)}, \ldots, n^{(G)})^\top$ and $\mathbf{1}$ is a vector of ones.
6) Sample the allocation indicators U_i independently for all subjects to create a new partition $\mathcal{P}$:
   a) Compute the full-conditioned classification probabilities $u_{i,g}(\boldsymbol{\theta};\ \boldsymbol{b}_i)$.

b) Sample new U_i from the multinomial distribution with probabilities $u_{i,g}(\boldsymbol{\theta};\, \boldsymbol{b}_i)$.

7) Sample e_0 using a Metropolis step from $p(e_0|\mathcal{P}, G) \propto p(\mathcal{P}|e_0, G) \cdot p(e_0)$.

This algorithm for model estimation has been implemented in R (R Core Team, 2022) with the use of the C programming language to optimise the computation time.

5.2 Post-processing

After omitting a suitable number of burn-in samples and applying thinning, the final MCMC chain contains M draws of $\boldsymbol{\theta}^m$, $\boldsymbol{U}^m$ and $\boldsymbol{b}_i^m$, $m = 1, \ldots, M$. For each draw m, the cluster indicators $\boldsymbol{U}^m$ induce cluster occupation numbers $\boldsymbol{n}^m = (n^{(1),m}, \ldots, n^{(G),m})^\top$ and a specific number of non-empty components $G_+^m = G - \sum_{g=1}^{G} \mathbb{1}(n^{(g),m} = 0)$. The number of non-empty components may differ among different draws m.

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

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

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

Before group-specific inference can be performed based on the MCMC samples, one potentially needs to resolve label switching (Redner and Walker, 1984). Because the likelihood as well as the prior and thus the posterior are label invariant, the posterior is multi-modal with modes corresponding to all parameterisations obtained by permuting the labels of unique components. The component labels may be switched across different draws of the MCMC sampler and a unique labelling needs to be obtained to determine an identified model where group-specific inference is possible. We suggest to use the procedure proposed in Frühwirth-Schnatter (2011) and Malsiner-Walli et al (2016) to resolve label switching with the later describing a method applicable when pursuing the sparse finite mixture approach.

In our simulation study and the applications, we observed that the number of filled components usually stabilises during MCMC sampling at a specific number, usually representing the lower bound of data clusters required to provide an adequate fit for the data. Initialising using a partition with all components being filled, we noted that during the first iterations of the MCMC algorithm superfluous components are emptied and only the necessary number of components required to represent the group structure in the data set

remains filled. The sparse finite mixture prior imposed on the component weights induces a penalty for the inclusion of redundant filled components, hence encouraging a solution where only a few components are filled. Monitoring thus the number of filled components serves as a means to assess convergence of the MCMC chain and thus decide on a suitable number of burn-in iterations to discard.

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

5.3 Classifying observations

After MCMC sampling there are basically two possibilities to obtain a final classification or partition of the subjects. The posterior classification probabilities $u_{i,g}$ may be estimated by conditioning not only on the observed data and parameter estimates, but also on estimates of the random effects $\boldsymbol{b}_i$. Given the MCMC samples this approach can easily be pursued for subjects included in the data set. We use this approach in the simulation studies (Section 6) as well as in the application using the EU-SILC data set (Section 8). This approach reduces the computational time needed because costly integral approximations are avoided and allows to obtain classifications based on the MCMC draws made for posterior inference anyway.

Alternatively the posterior classification probabilities $u_{i,g}$ can also be estimated by integrating out the random effects. This approach is applied in the second application considered (Section 7). It is computationally more expensive, but provides more accurate estimates because the latent random effects are not conditioned on, but integrated out.

Conditioning on random effects

The U_i^m draws obtained during MCMC sampling are posterior draws from the multinomial distribution with success probabilities equal to the a-posteriori probabilities induced by conditioning on the observed data as well as current parameter estimates and draws of the random effects. Their empirical means obtained with $\widehat{U}_{i,g} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}(U_i^m = g)$ represent suitable estimates for the classification probabilities taking into account uncertainty with respect to the parameter estimates as well as the random effects. This approach can be directly applied after the sampling procedure and post-processing and does not require any integral approximation. One only needs to store $n \cdot G \cdot M$ values. However, once the subjects are classified, these values may be discarded.

Based on these classification probabilities, subjects may be classified by assigning each subject i to the cluster g that has the highest estimate $\widehat{U}_{i,g}$ among all $g = 1, \ldots, G_+$. In case these classification probabilities are not clearly indicating assignment to a specific group, one may decide to leave those subjects unclassified. Rules to decide not to assign might for example be that the second largest classification probability lies within a pre-specified tolerance (such as 0.2) below the highest one or that the highest probability itself is below a given threshold (such as 0.6). Imposing such a rule leads to both classified subjects where classification is unambiguous, and unclassified subjects where assignment has been assessed to be not sufficiently clear.

Integrating out random effects

The posterior probabilities $u_{i,g}(\boldsymbol{\theta})$ are estimated for all sampled $\boldsymbol{\theta}^m$, i.e., the classifying probabilities are determined for the observed data and model parameters while integrating out the random effects. The posterior mean is then estimated by $\widehat{U}_{i,g} = \frac{1}{M} \sum_{m=1}^{M} u_{i,g}(\boldsymbol{\theta}^m)$. This approach requires $M \cdot G$ approximations of the integral, which is in particular costly when done for each subject $i = 1, \ldots, n$. Using the Laplacian approximation is in this case preferable to reduce the computational burden.

This approach approximates the posterior distribution of $u_{i,g}(\boldsymbol{\theta})$, thus allowing to construct 95% Highest Posterior Density (HPD) credible intervals. Subjects are then classified based on the highest $\widehat{U}_{i,g}$ value. Again for some subjects one may decide not to classify. In this case one can use as rule for example that the upper bound of the HPD intervals of the other groups need to lie below the lower bound of the HPD interval for group g to which one would assign based on the maximum value of $\widehat{U}_{i,g}$. This rule implies that one leaves a subject i unclassified if the classifying probability is comparable for more than one group and hence classification is not unambiguous.

6 Simulation study

We performed a simulation study to demonstrate the performance of our proposed approach under various settings. We were particularly interested in assessing how the structure of the sampled data as well as the data generating process affects (1) the ability to estimate the number of data clusters, (2) the clustering performance measured by the misclassification rates and (3) the accuracy of the model parameter estimates.

6.1 Simulation design

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

$n_i = 4$ observations per subject in order to mimic the panel structure of the applications.

For each data set we generate one outcome of each type – numeric Y^{N}, binary Y^{B}, ordinal Y^{O} with $K^{\mathsf{O}} = 5$ levels and general categorical Y^{C} with $K^{\mathsf{C}} = 4$ levels. With respect to the random-effects part, we only consider a random intercept term for each type of outcome $\boldsymbol{b}_i = (b_i^{\mathsf{N}}, b_i^{\mathsf{B}}, b_i^{\mathsf{O}}, b_i^{\mathsf{C}})^\top \sim \mathsf{N}_4(\mathbf{0}, \boldsymbol{\Sigma})$ and assume that the covariance matrix $\boldsymbol{\Sigma}$ of the random effects is the same across clusters and may be decomposed into standard deviations and correlation matrix such that

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

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

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

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

a) no effect of time at all (`no`),
b) a slope term common to all clusters (`parallel`),
c) different intercepts and slopes for each cluster resulting in a crossing (`cross`).

We follow the same scheme when specifying the models for estimation, considering models where no time effect is included, a common slope for time and a group-specific slope for time. Examples of the predictors simulated for the different time parameterisations and number of clusters G are illustrated in Figure 1.

The intercept term is always (both when generating the data set and when estimating) considered to be group-specific. This ensures some differences between clusters. The numerical outcome is obtained by adding an error term with group-specific standard deviation, $\{0.5, 0.8\}$ for $G = 2$ and $\{0.5, 0.75, 1\}$

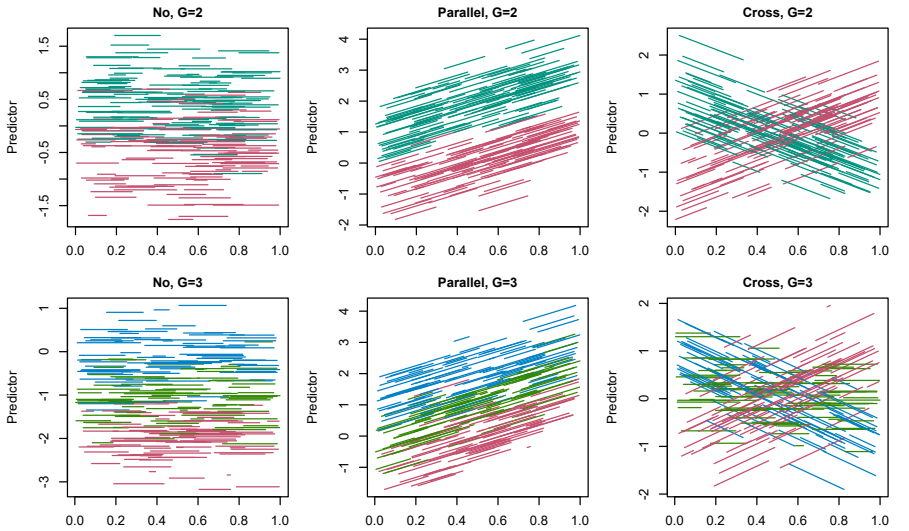

Fig. 1 Lines connecting predictors of $n = 250$ individual subjects generated from G clusters for different types of time effects. The maximum length of the observational window is $\xi = \frac{1}{3}$

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

We generate 200 data sets for each considered data setting. For Bayesian inference, the prior distributions together with their parameter values are specified as outlined in Section 4. For estimating the number of data clusters or assessing the clustering abilities, we initialise the Markov chain with the maximal number of components $G_{\mathsf{max}} = 10$ considered for the mixture model. A burn-in period of $B = 500$ samples was enough to then use the next $M = 10\,000$ sampled parameter and latent variable values to approximate their posterior distributions. Subjects were classified using the sampled indicators U_i, leaving subjects unclassified when less than 60% of these indicators assigned the subject to the same cluster.

6.2 Estimating number of data clusters and classifying subjects

In the following we assess the ability of the proposed approach to estimate the number of data clusters and evaluate the classification performance, focusing in particular on the benefit incurred through joint modelling of the outcome variables. We consider the `cross` parameterisation of time with $\xi = \frac{1}{3}$ for data generation and also use a suitable model specification for estimation to be

able to capture these effects. We estimate the model for each type of outcome separately as well as all four outcomes of different types jointly.

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

Figure 2 provides an overview on the proportions of correctly classified, unclassified and misclassified subjects when using either only a single outcome variable or using all four outcome variables jointly. In addition the sample size n and the true number of data clusters are also varied. The results for the single outcome variables are shown in the rows labelled "Num" for numeric outcome, "Bin" for binary outcome, "Ord" for ordinal outcome and "Cat" for general categorical outcome. The results when modelling all four outcomes jointly are shown on top in the row labelled "All".

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

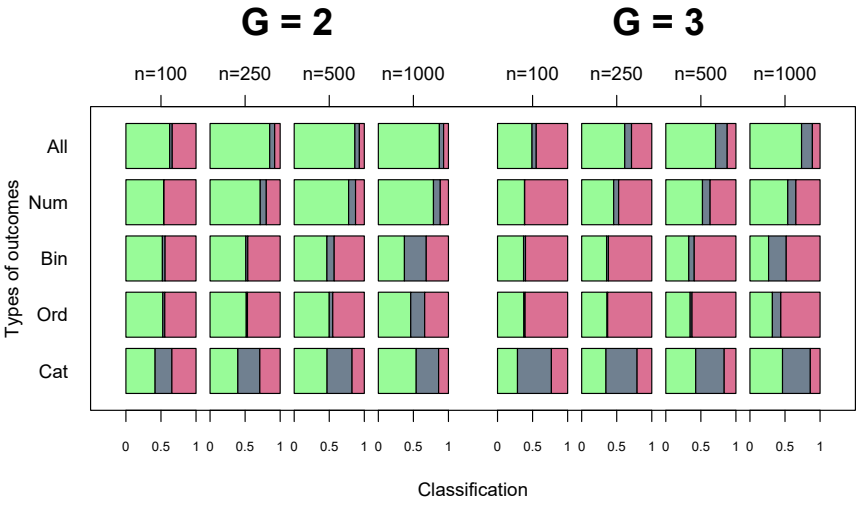

Fig. 2 Proportions of correctly classified (green), unclassified (grey) and misclassified (red) subjects in dependence of the types of outcomes used, sample size n and the true number of data clusters G. The number of data clusters used for classification are estimated based on $\widehat{G}_+$, the most frequent number of non-empty components during MCMC sampling with $G_{\text{max}} = 10$

Figure 2 also highlights the impact of the type of outcome on the classification performance. If only a single outcome is considered, the numeric outcome performs best, while the categorical outcome results in a classification performance which is barely better than a random classification. Modelling all types together clearly outperforms the single models and achieves the highest correct classification rates indicating the advantage of using a modelling approach which allows to jointly model the data.

6.3 Estimating model parameters

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

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

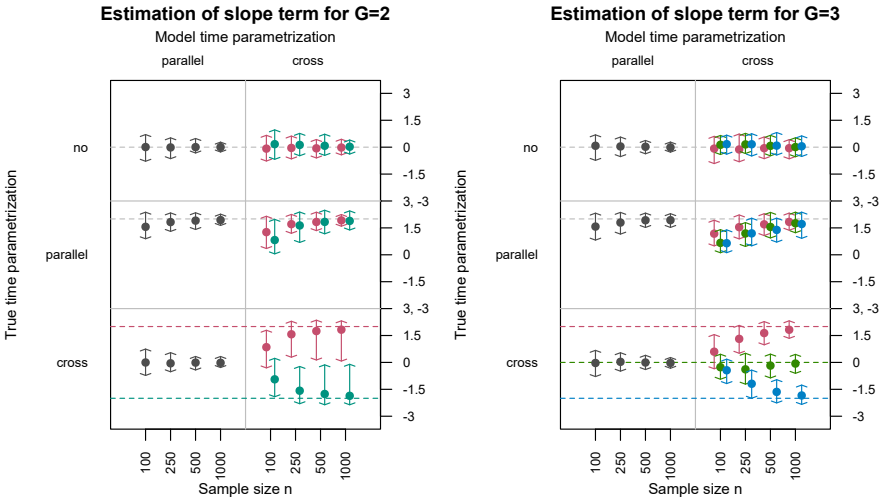

Fig. 3 Medians, 2.5 and 97.5% quantiles of estimated posterior medians of the slope term for the Bin outcome variable across 200 simulated data sets. Model estimation is performed assuming that the number of data clusters G is known. Different settings are considered for the effect of the time covariate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

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

7 Analysis of the PBC medical study data

In the study of primary biliary cholangitis (PBC) of liver conducted by the Mayo Clinic between 1974 and 1984, 312 patients were randomly assigned to a placebo control group and to a treatment group consisting of D-penicillamine drug users. The study protocol required visits after 6 months, one year and then annually until the patients died, had a liver transplant or dropped out from the study. At each visit multiple laboratory results were obtained and combined into a longitudinal data set. At each visit not all tests were undertaken leading to missing values in the outcome variable.

This data set has in particular been studied to predict survival, see Therneau and Grambsch (2000). In the following we use the data to infer different prognosis groups based on the observed patterns of evolvement of specific markers over time taking also age and gender into account as covariates. Having established an association of the groups identified with survival, new patients may be classified based on their marker evolvement.

7.1 Data and model description

Similar to Komárek and Komárková (2013), we restrict our analysis to the patients ($n = 260$) who survived the first 910 days (2.5 years) of the study without liver transplantation. The vast majority (178) of patients have $n_i = 4$ visits recorded within this period. However, there are also patients included where only a single visit is available. Restricting the data to only the first 910 days imitates a situation, where a prognosis for a patient is desired and the aim is to establish a classification rule for patients where data from 910 days of the follow-up are available.

We used five outcome variables for the analysis. Two numeric markers are included as outcome variables: serum bilirubin (*bili*) and *albumin* (on log-scale). Two binary outcome variables are included which indicate if the patient

suffered from presence of blood vessel malformations in the skin (*spiders*) and hepatomegaly or enlarged liver (*hepato*). A single ordinal outcome variable is included which indicates the seriousness of *edema*. Missing values were augmented during MCMC sampling to keep all subjects in the analysis and obtain a posterior approximation of the unknown values.

The five markers are jointly modelled by assuming random intercepts for the patients and a group-specific linear effect of time, age at entry to the study and gender without any interaction terms. All other model parameters were also considered to be group-specific to capture differences in all possible aspects. Hence, not only a different evolution over time is expected, but also the effects of age or gender may vary across groups as well as the noise variances for the numeric outcomes and the covariance structure of the random intercepts.

A sparse finite mixture was induced by setting $a_e = 1, b_e = 100$ and the other hyperparameters were set to correspond to a unit scale prior distribution. With $G_{\mathsf{max}} = 10$ the MCMC sampling converged after few hundred steps to a $\widehat{G}_+ = 2$ solution. The burn-in period was decided as a multiple of 200 iterations based on a visual inspection of trace plots. For the results reported, $M = 10\,000$ sampled parameter and latent variable values were used without thinning to approximate their posterior distributions. Repeating this procedure for four different chains using random initialisations indicated that results are rather comparable across the chains.

7.2 Results

The $n = 260$ patients are classified based on the maximum classifying probabilities obtained by integrating out the random effects. This results in a partition of the patients into two groups. Combining this grouping with the remaining data (beyond 910 days) allows to determine the Kaplan–Meier estimates of the survival functions for each group (see Figure 4). Even though the fitted model did not include the information on subsequent survival, the identified groups clearly exhibit different survival curves. Thus the grouping identified can be used to obtain prognosis about future survival.

Table 1 provides posterior estimates of the group-specific parameters which allow to characterise the two estimated groups with respect to their covariate effects. The red cluster of Figure 4 ($n^{(1)} = 106$, 26 men) with the drastically decreasing survival function represents about 41% of patients with high serum bilirubin increasing over time, lower serum albumin in general, increasing odds of spiders with time and increasing risk of edema over time. On the other hand, the turquoise cluster ($n^{(2)} = 154$, one man only) with much higher survival probabilities consists of 59% of the patients who have a low value of serum bilirubin only slowly increasing over time, higher values of serum albumin, stable odds in time for both spiders and hepatomegaly and increasing risk of edema with age.

Table 1 Posterior medians of group-specific model parameters including 95% equal-tailed credible intervals

Type	Outcome	Cluster	β_0 or c_k	β_{time}	$10\beta_{\text{age}}$	β_{sex}	$\tau^{-\frac{1}{2}}$
Numeric	log(bili)	1	0.85 (0.18; 1.59)	0.21 (0.14; 0.28)	−0.02 (−0.15; 0.12)	0.10 (−0.25; 0.51)	0.50 (0.46; 0.55)
		2	0.14 (−0.24; 0.53)	0.02 (−0.02; 0.05)	−0.09 (−0.16; −0.02)	0.20 (−0.16; 0.56)	0.24 (0.23; 0.26)
	log(albumin)	1	1.31 (1.19; 1.42)	−0.02 (−0.04; 0.00)	−0.01 (−0.03; 0.01)	−0.03 (−0.08; 0.02)	0.16 (0.15; 0.18)
		2	0.80 (0.69; 0.91)	−0.01 (−0.02; 0.00)	0.01 (0.00; 0.02)	0.45 (0.35; 0.55)	0.12 (0.11; 0.13)
Binary	spiders	1	−0.45 (−2.19; 1.26)	0.50 (0.09; 0.93)	−0.30 (−0.68; 0.07)	0.55 (−0.64; 1.73)	
		2	−0.11 (−1.84; 1.63)	0.08 (−0.39; 0.54)	−0.63 (−1.15; −0.18)	−0.19 (−1.85; 1.51)	
	hepato	1	−0.17 (−1.91; 1.62)	0.32 (−0.12; 0.78)	0.30 (−0.09; 0.69)	−0.10 (−1.39; 1.14)	
		2	0.24 (−1.42; 1.92)	0.07 (−0.27; 0.41)	−0.21 (−0.59; 0.13)	−0.45 (−2.00; 1.13)	
Ordinal	edema	1	4.09 (2.78; 5.50) 7.05 (5.47; 8.80)	1.07 (0.63; 1.53)	0.56 (−0.07; 1.20)	0.62 (−0.65; 1.90)	
		2	1.63 (−0.13; 3.86) 6.75 (4.33; 10.06)	0.08 (−0.39; 0.53)	0.82 (0.07; 1.68)	−2.26 (−4.05; −0.46)	

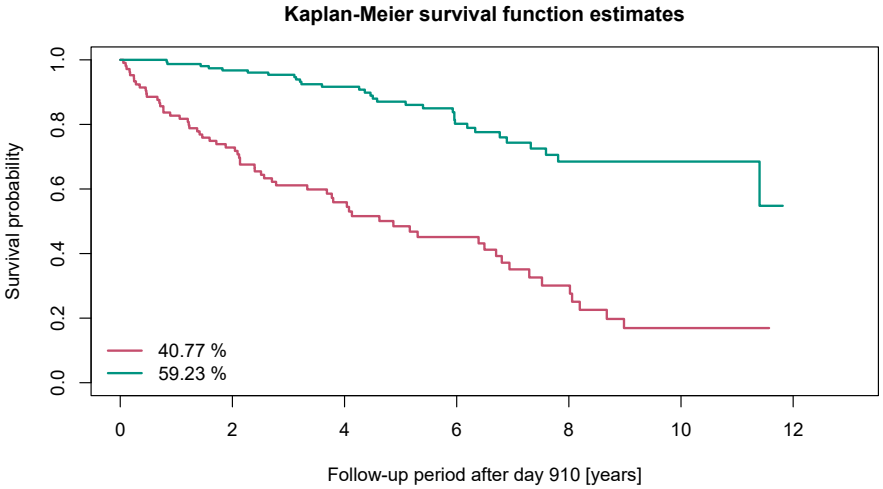

Fig. 4 Kaplan–Meier survival function estimates after day 910 of the $n = 260$ patients from the PBC medical study clustered into the two estimated groups

8 Analysis of the EU-SILC data

The EU-SILC (Statistics on Income and Living Conditions) survey gathers data on households within member states of the European Union, Iceland, Norway and Switzerland annually since 2003. We apply our proposed approach to identify groups of households which differ in their evolvement of financial capability over time as measured by several highly correlated outcomes of mixed type.

8.1 Data and model description

The analysis focuses on the subset of Czech households surveyed between 2005 and 2018. This time period includes the years of the economic crisis which started in late 2008. We have $n = 23\,360$ households that were followed for exactly $n_i = 4$ consecutive years, as induced by the rotational design of the study. Starting with more than 7 000 households, each year a quarter is dropped to be replaced by a comparably sized set of new households.

Eight outcomes (two for each type) are modelled jointly using the proposed approach. All eight outcomes reflect the financial capacity of the household. Two numeric outcome variables are included which are income related: *Equivalised total disposable income* [€/year], that sums the gross personal income components of all household members over the whole year and divides it by the *Equivalised household size* (see below), and *Lowest monthly income to make ends meet* [€/month], that reflects the minimum net monthly income required to pay for all usual necessary expenses of the household. In addition, the financial capacity is measured by the ability to afford certain luxuries. *Affordability*

of one week annual holiday away from home and *Capacity to face unexpected financial expenses* are binary outcome variables ("Yes", "No"), while the possession indicators of a car or a computer are general categorical outcome variables with three levels consisting of "Yes", "No – cannot afford" and "No – other reason". Two ordinal outcome variables are also included which rather reflect subjective assessment of financial capability and are measured as the *Ability to make ends meet* (on a scale from 1="with great difficulty" to 6="very easily") and the perceived *Financial burden of the total housing cost* (with levels: "a heavy burden", "a slight burden", "not a burden at all").

Because the numeric income related outcomes have a heavily skewed distribution, we transformed the values to log-scale. In case the income was negative (which very rarely occurred), it was set to zero on the log-scale. The baseline levels for the general categorical and the ordinal outcomes were determined by ordering the categories with respect to their expected positive correlation with increasing financial capacity.

In the regression the time variable indicating the year when the survey was completed was included as group-specific covariate to identify how the financial capacity of the households evolves over time, in particular also during the phasis of an economic crises. To capture a possible change in trend, we used a quadratic spline parameterisation with one inner knot.

Additional covariates were also included in the regression which characterise the households. These covariates were included with constant effects across the whole population. These additional variables are: *Level of urbanisation* of their location (with levels "thinly-populated area", "intermediate area", "densely populated area", "capital city of Prague"), the *Highest education* level attained by at least one household member (with levels "lower than secondary", "secondary", "higher than secondary") whether at least one household member is a baby (i.e., younger than 3 years) or a student (i.e., attending some educational institution) and the *Equivalised household size*. The *Equivalised household size* is obtained by summing over all household members using the following weights: a weight of 1 for the first member, a weight of 0.5 for the other household members older than 14 and a weight of 0.3 for household members who are 14 or younger.

The maximum number of components was set to $G_{\mathsf{max}} = 20$. To invoke sparsity we specify the parameters of the prior distribution for e_0 to be $a_e = 1$, $b_e = 100$. To regularise the effect estimates and shrink them towards zero, the standard deviations of the priors for the centred effects were set to 0.5. Ordered intercepts $\boldsymbol{c}$ and error term precisions $\boldsymbol{\tau}$ are set to be group-specific, the variance matrix $\boldsymbol{\Sigma}$ of random effects is kept common to all households.

The burn-in period was decided as a multiple of 1 000 iterations based on a visual inspection of trace plots. For the results reported, $M = 1\,000$ sampled parameter and latent variable values were used without thinning to approximate their posterior distributions. Repeating this procedure for four different chains using random initialisations indicated again comparability of results obtained across chains.

8.2 Results

Post-processing led to the estimation of $\widehat{G}_+ = 4$ clusters. Using the sampled U_i, we classified households where the maximum classification probability was at least 0.5. Otherwise the household remained unclassified (0.55% of households). The classified households were used to create the plots in Figure 5 describing the evolvement of the outcome variables across time for each cluster separately.

Figure 5 indicates that the cluster sizes vary strongly with the green cluster containing 78.57% of the households, followed by the yellow cluster consisting of 17.11% of the households. With respect to their cluster size, the remaining two clusters seem rather negligible with the blue cluster containing 3.65% of the households and the red cluster containing 0.12% of the households. 0.55% of the households remained unclassified.

Assessing the financial capacity of the households as depicted by the outcome variables, one can conclude that the blue cluster (3.65%) contains households that are doing well in general, while the yellow cluster (17.11%) represents the more struggling households. In-between these two, the most common green cluster (78.57%) contains housholds with an intermediate financial capacity. The red cluster (0.12%) consists of the rare households faced with a bad financial situation.

The group-specific evolvement of the log-scaled *Equivalised total disposable income* is shown in Figure 5 on the top left. In particular the estimated posterior median curves indicate how the evolvement differs across the clusters. For all four clusters a rather strong increase is captured for the first four years which at the start of the economic crisis either levels off to a rather constant equivalised total disposal income or even to a slightly decreasing one.

The plot of the *Equivalised total disposable income* also indicates that the clusters strongly differ in the standard deviation of the error term in the linear regression model. This standard deviation captures how much the income variable differs for the same household across the four consecutive measurements and thus also reflects how volatile the income situation is for a houshold. Assessing the group-specific standard deviation estimates for the error term using 95% equi-tailed credible intervals indicates that the green cluster contains households with a rather constant income over time ($\widehat{\sigma} \in (0.063;\ 0.064)$). On the contrary, the yellow ($\widehat{\sigma} \in (0.129;\ 0.133)$) and, especially, the blue cluster ($\widehat{\sigma} \in (0.279;\ 0.300)$) contain households with dramatic changes in their income between consecutive years. The small red cluster contains the extremely low values of income (including the few negative income observations) that also heavily fluctuate from one year to another.

The posterior estimates for the other covariates included in the regression with a constant effect for the whole population indicate, based on the posterior medians and the 95% equi-tailed credible intervals, that living in more densely populated areas (town, city, Prague) increases the expected *Equivalised total disposable income* by 1.27% (0.93%; 1.61%), 2.12% (1.68%; 2.48%), 8.29% (7.63%; 8.98%), respectively, compared to living in a thinly populated village. Having one additional adult within a household increases the expected

Equivalised total disposable income by 2.58% (2.41%; 2.73%). Taking care of a baby, respectively having a student, within a household decreases the expected *Equivalised total disposable income* by 5.44% (5.08%; 5.76%), respectively 3.30% (3.02%; 3.56%). Having as highest education level a secondary, respectively upper-secondary or tertiary education level within a household

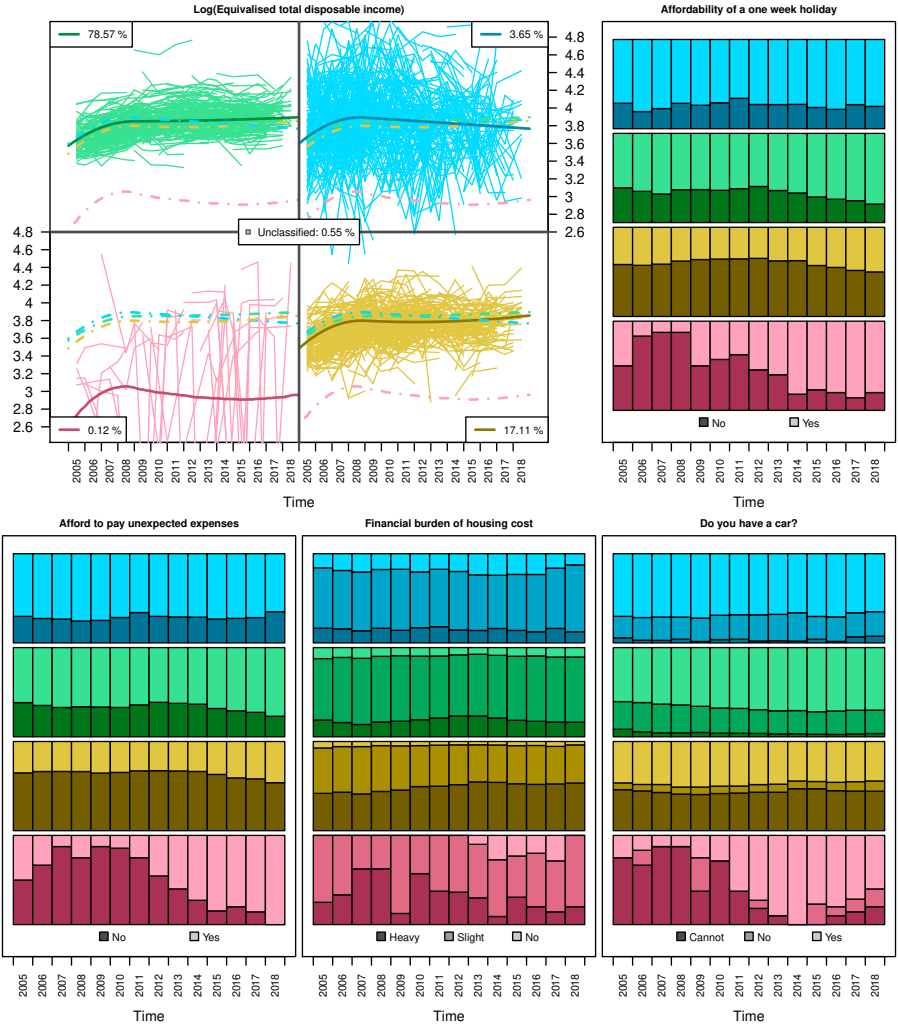

Fig. 5 Visualisation of the evolution of five (out of the eight) outcome variables across time when grouped into $\widehat{G}_+ = 4$ clusters. The upper left plot shows the observed values for the classified households together with the estimated median posterior curves for the log-transformed numeric outcome variable *Equivalised total disposable income* for a representative household of size 1.5 from a thinly-populated area, with secondary educational level the highest achieved level of all household members and without having a baby or a student as household member. The other plots visualise the empirical frequencies of the categorical outcome variables after classifying households obtained separately for each year

increases the expected *Equivalised total disposable income* by 11.02% (10.61%; 11.37%), respectively 20.44% (19.85%; 21.08%), compared to a situation when the highest education level achieved by all household members corresponds to the lower-secondary educational level.

The estimated curves for the categorical outcome variables across time were more or less flat for all four clusters, but they differed in their levels across clusters. This agrees with Figure 5 where we barely see any evolution of the ratios in time within any of the main clusters (blue, green, yellow). These constant ratios, however, correspond to the interpretation of the clusters obtained so far. Households within the blue cluster most probably own a car, can afford a week holiday away from home, have capacity to pay for unexpected expenses and the housing cost does not seem to be a burden to them when compared to other clusters. On the other hand, the majority of households within the yellow cluster cannot afford a car, nor a week holiday, nor pay for unexpected expenses. They also agree to the highest extent with the statement that housing cost is a heavy financial burden. Households within the green cluster (the majority) are comparable to the prosperous ones in the blue cluster, but they are in general a bit worse off. We obtained an analogous interpretation for the other outcomes which are not included in Figure 5 and these results are hence not shown.

9 Conclusion

This paper proposes an approach which allows to infer clusters from multivariate longitudinal data of possibly different types by building on and combining several different methodologies. The model specification allows to combine an arbitrary number of numeric, binary, ordinal or general categorical outcome variables and to model them jointly by GLMMs. The suitable distributional family is used for each outcome type and the linear predictor may consist of group-specific as well as common fixed effects as well as group-specific or common random effects. The random effects are assumed to follow a multivariate normal distribution with a general covariance matrix and are allowed to be correlated not only within a single outcome but also across all outcome variables. This accounts for correlation between observations from the same subject even after accounting for group differences and any covariate effects in the regression. A finite mixture model is specified to embed the clusterwise regression problem into a model-based clustering framework.

The Bayesian approach is pursued for model estimation and inference exploiting the possibility to determine the number of data clusters based on a sparse finite mixture approach, specify priors which have a regularising effect on the mixture likelihood and fully exploit the hierarchical structure and the latent variable framework using Bayesian data augmentation in MCMC inference.

The performance of the proposed approach is evaluated in a simulation study indicating the benefits of jointly modelling the outcome variables to

improve the clustering abilities as well as highlighting the accuracy of the parameter estimates obtained from an identified mixture model. The applications demonstrate how the proposed approach helps analysing medical and economic survey data indicating the wide potential in many different areas such as health care, psychology, social sciences and many more.

Acknowledgments. This research was supported by the Czech Science Foundation (GAČR) grant 19-00015S, the Charles University, project GA UK No. 298120, and the Austrian Science Fund (FWF) grant P28740.

Appendix A Full-conditioned distributions in Gibbs sampling

In this section we list the full-conditioned distributions for the model parameters which fall into well-known distributional families. The rest of the parameters is sampled using a Metropolis proposal step and their detailed derivations are postponed to Appendix B. Note that in the following derivations the parameters are considered to be group-specific where applicable. Similar formulas may be derived even if some of them were kept common to all clusters.

A.1 Component sizes $\boldsymbol{w}$

Due to conjugacy the full-conditioned distribution of $\boldsymbol{w}$ stays within the family of Dirichlet distributions, i.e., $\boldsymbol{w}|\boldsymbol{U}, e_0 \sim \mathsf{Dir}_G\left(\boldsymbol{n}(\boldsymbol{U}) + e_0\mathbf{1}\right)$. The Dirichlet parameter of the full-conditioned posterior distributed is obtained by adding to the prior value e_0 the number of subjects currently assigned to each of the clusters, i.e., $\boldsymbol{n}^G(\boldsymbol{U}) = \{n^{(g)}(\boldsymbol{U}) = \sum_{i=1}^{n} \mathbb{1}\{U_i = g\};\ g = 1, \ldots, G\}$.

A.2 Group-allocation indicators $\boldsymbol{U_i}$

The latent variables U_i are discrete variables taking values in $\{1, \ldots, G\}$. To draw their values from the multinomial distribution, we use Bayes' theorem to calculate the full-conditioned probability that the i-th subject is from group g:

$$u_{i,g}(\boldsymbol{\theta};\ \boldsymbol{b}_i) = \frac{w_g\, p\left(\mathbb{Y}_i \,\middle|\, \boldsymbol{b}_i, U_i = g, \boldsymbol{\beta}^{(g)}, \boldsymbol{\Sigma}^{(g)}, \boldsymbol{\tau}^{(g)}, \boldsymbol{c}^{(g)};\ \mathcal{C}_i\right) p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\Sigma}^{(g)}\right)}{\sum_{g'=1}^{G} w_{g'}\, p\left(\mathbb{Y}_i \,\middle|\, \boldsymbol{b}_i, U_i = g', \boldsymbol{\beta}^{(g')}, \boldsymbol{\Sigma}^{(g')}, \boldsymbol{\tau}^{(g')}, \boldsymbol{c}^{(g')};\ \mathcal{C}_i\right) p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\Sigma}^{(g')}\right)}. \quad \text{(A1)}$$

By conditioning also on the random effects, we work with the contributions to the likelihood as given in Section 2.1 without the necessity of integration as in (5).

A.3 Precision parameter e_0

We follow Malsiner-Walli et al (2016) and sample e_0 from the semi-marginalised full-conditioned distribution. We integrate the parameter $\boldsymbol{w}$ out of the conditioning and sample e_0 from $p(e_0|\boldsymbol{U}) \propto p(e_0)p(\boldsymbol{U}|e_0)$ instead of $p(e_0|\boldsymbol{U}, \boldsymbol{w})$. The integration over $\boldsymbol{w}$ yields

$$p(\boldsymbol{U}|e_0) = \frac{\Gamma(Ge_0)}{\Gamma(Ge_0+n)} \prod_{\{g:n^{(g)}(\boldsymbol{U})>0\}} \frac{\Gamma(n^{(g)}(\boldsymbol{U})+e_0)}{\Gamma(e_0)}. \tag{A2}$$

Combining (A2) with the pdf (10) of the prior for e_0, we obtain

$$p(e_0|\boldsymbol{U}) \propto e_0^{a_e-1}\exp\{-b_e e_0\} \cdot \frac{\Gamma(Ge_0)}{\Gamma(Ge_0+n)} \prod_{\{g:n^{(g)}(\boldsymbol{U})>0\}} \frac{\Gamma(n^{(g)}(\boldsymbol{U})+e_0)}{\Gamma(e_0)}. \tag{A3}$$

We are not able to sample from this distribution directly. However, we can still use a Metropolis step, see appendix section B.1. Since e_0 is restricted to be positive, we perform the proposal on the log-scale by defining a new parameter $e_0^\star = \log e_0 \in \mathbb{R}$, which yields

$$p(e_0^\star|\boldsymbol{U}) \propto \exp\{e_0^\star a_e - b_e\exp\{e_0^\star\}\} \cdot \frac{\Gamma(G\exp\{e_0^\star\})}{\Gamma(G\exp\{e_0^\star\}+n)} \prod_{\{g:n^{(g)}(\boldsymbol{U})>0\}} \frac{\Gamma(n^{(g)}(\boldsymbol{U})+\exp\{e_0^\star\})}{\Gamma(\exp\{e_0^\star\})}. \tag{A4}$$

A.4 Parameters for numeric outcomes

The assumption of normality for numeric outcomes and the choice of the semi-conjugate prior distributions for $\boldsymbol{\tau}$ and $\boldsymbol{\beta}_r, r \in \mathcal{R}^{\mathsf{Num}}$ preserves the distributional families for the full-conditioned distributions. The inverse variance $\tau_r^{(g)}$ for the numeric outcome $r \in \mathcal{R}^{\mathsf{Num}}$ within cluster $g = 1, \ldots, G$ follows a Gamma distribution

$$\tau_r^{(g)} \,\Big|\, \boldsymbol{Y}^r, \boldsymbol{U}, \boldsymbol{b}^r, \boldsymbol{\beta}_r^{(g)};\ \mathcal{C}^r \ \sim\ \Gamma\left(\alpha_{r,1}^{(g)}, \alpha_{r,2}^{(g)}\right)$$

with updated parameters $\alpha_{r,1}^{(g)}$ and $\alpha_{r,2}^{(g)}$:

$$\alpha_{r,1}^{(g)} = \frac{1}{2}\sum_{\{i:U_i=g\}} n_i + \frac{1}{2}d_r^{\mathsf{F}} + \alpha_1,$$

$$\alpha_{r,2}^{(g)} = \frac{1}{2}\sum_{\{i:U_i=g\}}\sum_{j=1}^{n_i}\left(Y_{i,j}^r - \eta_{i,j}^{r,(g)}\right)^2 + \frac{1}{2}\sum_{j=1}^{d_r^{\mathsf{F}}}\frac{\left(\beta_{r,j}^{(g)} - \beta_{0,r,j}^{(g)}\right)^2}{d_{j,j}^r} + \alpha_2.$$

If $\tau_r \equiv \tau$ were common to all clusters, we would use all subjects $i = 1, \ldots, n$ instead of just the subjects assigned to cluster g.

The full-conditioned distribution for the fixed effects $\boldsymbol{\beta}_r^{(g)}$ of the numeric outcome $r \in \mathcal{R}^{\mathsf{Num}}$ is a multivariate normal distribution with

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

where $\mathcal{N}_g(\boldsymbol{U}) = \{i : U_i = g\}$ and

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

$\mathbb{D}^r = \mathsf{diag}\{d^r_{j,j}, j = 1, \ldots, d_r^{\mathsf{F}}\}$ is the diagonal variance matrix of the prior distribution and $\bullet_{\mathcal{N}_g(\boldsymbol{U})}$ restricts $\bullet$ to the subset of subjects in group g:

$$\mathbb{X}^r_{\mathcal{N}_g(\boldsymbol{U})} = \begin{pmatrix} \vdots \\ \left(\boldsymbol{x}^r_{i,j}\right)^\top \\ \vdots \end{pmatrix}_{\substack{i\in\mathcal{N}_g(\boldsymbol{U}),\\ j=1,\ldots,n_i}}, \qquad \widetilde{\boldsymbol{y}}^r_{\mathcal{N}_g(\boldsymbol{U})} = \begin{pmatrix} \vdots \\ Y^r_{i,j} - \eta^{\mathsf{R},r}_{i,j} \\ \vdots \end{pmatrix}_{\substack{i\in\mathcal{N}_g(\boldsymbol{U}),\\ j=1,\ldots,n_i}}.$$

In case some of the fixed effects $\boldsymbol{\beta}_r$ are common to all groups, the linear combination $\left(\boldsymbol{x}^r_{i,j}\right)^\top \boldsymbol{\beta}_r^{(g)}$ is divided into a sum of two linear combinations of lower dimension. Then the evaluation of the full-conditioned distribution and subsequent sampling is performed separately for the part which is common to all groups and the group-specific part. In each of these separate steps the other part is simply subtracted to obtain the auxiliary vector $\widetilde{\boldsymbol{y}}$. Unlike the group-specific part, where we work with the subjects currently belonging to the g-th cluster only, the full-conditioned distribution of the common fixed part uses all subjects.

A.5 Prior scale matrix $\mathbb{Q}$ for $\boldsymbol{\Sigma}$

Parameter $\mathbb{Q}$ is the hyperparameter to increase the flexibility of the prior distribution of $\boldsymbol{\Sigma}$. Priors for both, $\boldsymbol{\Sigma}$ and $\mathbb{Q}$, are specified for their inverse counterparts because these are more natural to work with. The specification implies that the Wishart distribution family for the inverse matrix $\mathbb{Q}^{-1}$ is

preserved with parameters

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

A.6 Prior inverse covariance matrices $\boldsymbol{\Sigma}^{-1}$ for random effects $\boldsymbol{b}$

Similarly, the multivariate normal assumption for the random effects $\boldsymbol{b}_i$ with variance matrix $\boldsymbol{\Sigma}^{(g)}$ keeps the full-conditioned distribution of the precision matrix $\boldsymbol{\Sigma}^{-(g)}$ within the family of Wishart distributions

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

independently for all $g = 1, \ldots, G$, where

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

If $\boldsymbol{\Sigma}^{(g)} \equiv \boldsymbol{\Sigma}$ is common to all clusters, then we would use the random effects from all subjects $i = 1, \ldots, n$.

Appendix B Metropolis proposal step with the use of the Newton–Raphson method

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

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

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

$$\alpha\left(\boldsymbol{\omega}^{m+1}, \boldsymbol{\omega}^m\right) = \min\left\{1, \frac{p\left(\boldsymbol{\omega}^{m+1}\right)}{p\left(\boldsymbol{\omega}^m\right)}\right\}$$

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

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

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

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

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

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

To find $\widehat{\boldsymbol{\omega}}$ maximising $\ell(\boldsymbol{\omega})$, we employ the Newton–Raphson method. Starting from some initial value $\boldsymbol{\omega}_0$, e.g., the maximum from the previous step, we iterate the following steps until convergence:

a) Evaluate the gradient and Hessian matrix of $\ell(\boldsymbol{\omega})$ at the current value $\boldsymbol{\omega}_k$.
b) Use the Cholesky decomposition to solve the following system of equations:
$$\left[-\frac{\partial^2 \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^{\top}}\bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_k} \right] \boldsymbol{s} = \frac{\partial \ell(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_k}.$$
c) Use the solution $\boldsymbol{s}$ to define a new value $\boldsymbol{\omega}_{k+1} = \boldsymbol{\omega}_k + \boldsymbol{s}$.
d) Check the convergence by computing the norm $\|\boldsymbol{s}\|$ of the step $\boldsymbol{s}$ and continue if still too large.

This procedure yields $\widehat{\boldsymbol{\omega}}$ and the basis for the precision matrix $\boldsymbol{\Omega}^{-1}$ of the incremental distribution.

In the following sections we explore in detail the peculiarities of individual parameters that require a Metropolis proposal approach for sampling from the full-conditioned distribution. These include: the log-precision $e_0^{\star}$ (to sample $e_0 > 0$), the fixed effects $\boldsymbol{\beta}_r^{(g)}$ of categorical outcomes $r \in \mathcal{R} \backslash \mathcal{R}^{\mathsf{Num}}$, the random effects $\boldsymbol{b}_i$ specific to each subject i and the transformed ordered intercepts

$\boldsymbol{a}_r^{(g)}$ (to sample the ordered intercepts $\boldsymbol{c}_r^{(g)}$ and the corresponding probabilities $\boldsymbol{\pi}_r^{(g)}$).

B.1 Log-precision parameter $e_0^\star$

We start with the only univariate parameter – the precision $e_0 > 0$, which has to be transformed into $e_0^\star = \log e_0$ such that it has as domain the whole $\mathbb{R}$. Equation (A4) transformed into log-scale yields

$$\begin{aligned}\ell(e_0^\star|\,\boldsymbol{U}) = \;& \text{const.} + e_0^\star a_e - b_e \exp\{e_0^\star\} + \log\Gamma(G\exp\{e_0^\star\}) \\ & - \log\Gamma(G\exp\{e_0^\star\} + n) + \sum_{\{g:n^{(g)}(\boldsymbol{U})>0\}} \log\Gamma(n^{(g)}(\boldsymbol{U}) + \exp\{e_0^\star\}) \\ & - G_+ \log\Gamma(\exp\{e_0^\star\}). \quad \text{(B5)}\end{aligned}$$

The first and second derivative of (B5) can be obtained with the use of the derivatives of the log-Gamma function $\log\Gamma$, namely the digamma function ψ and the trigamma function ψ_1, both implemented in base R. They take the following form:

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

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

B.2 Fixed effects $\boldsymbol{\beta}_r$ for categorical outcomes

Table B1 contains an overview of the contributions of a single outcome observation to the log-likelihood depending on the type of the outcome. Moreover, derivatives with respect to the predictor η (or $\boldsymbol{\eta}$) can be further used for determining the derivatives with respect to fixed and random effects. In this section,

which is devoted to the fixed effects, we will use that

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

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

In the following we present the log-posteriors and their derivatives for the full-conditioned distribution of the fixed effects $\boldsymbol{\beta}_r^{(g)}$ within the g-th group for a binary, ordinal and general categorical outcome. We start with a binary outcome, $r \in \mathcal{R}^{\mathsf{Bin}}$:

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

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

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

Table B1 The contribution of a single observation to the log-likelihood as well as the first and second derivative depending on the type of the outcome. Formulas for ordinal and categorical outcomes assume $Y = k$. Categorical outcomes have a multivariate predictor $\boldsymbol{\eta}$, the other types work with a univariate predictor η. Notation follows the one used in Section 2.1

Type	$\ell(Y \mid \boldsymbol{\eta}, \boldsymbol{\zeta})$	$\frac{\partial}{\partial \boldsymbol{\eta}} \ell(Y \mid \boldsymbol{\eta}, \boldsymbol{\zeta})$	$-\frac{\partial^2}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top} \ell(Y \mid \boldsymbol{\eta}, \boldsymbol{\zeta})$
Num	$-\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\tau - \frac{\tau}{2}(Y-\eta)^2$	$\tau(Y-\eta)$	τ
Bin	$Y\eta - \log\left(1+\exp\{\eta\}\right)$	$Y - \mathsf{logit}^{-1}(\eta)$	$\mathsf{logit}^{-1}(\eta)\left(1-\mathsf{logit}^{-1}(\eta)\right)$
Ord	$\log(q_k) = \log(p_{k-1} - p_k)$	$1 - p_{k-1} - p_k$	$p_{k-1}(1-p_{k-1}) + p_k(1-p_k)$
Cat	$\eta_k - \log\left(1 + \sum_{k'=1}^{K-1} \exp\{\eta_{k'}\}\right)$	$\boldsymbol{e}_k - \mathsf{softmax}(\boldsymbol{\eta}):$ if $k < K$ $-\mathsf{softmax}(\boldsymbol{\eta}):$ if $k = K$	$\mathsf{diag}\{\mathsf{softmax}(\boldsymbol{\eta})\} -$ $\mathsf{softmax}(\boldsymbol{\eta})\,\mathsf{softmax}(\boldsymbol{\eta})^\top$

$$+p_{Y_{i,j}^r}\left(1-p_{Y_{i,j}^r}\right)\Big]\,\boldsymbol{x}_{i,j}^r\left(\boldsymbol{x}_{i,j}^r\right)^\top+\mathbb{D}_r^{-1}.$$

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

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

The first derivative consists of the following subvectors:

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

The negative Hessian matrix consists of the following blocks:

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

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

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

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

B.3 Random effects $\boldsymbol{b_i}$

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

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

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

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

in combination with the derivatives in Table B1 to compute the derivatives of full-conditioned log-posterior of $\boldsymbol{b}_i$ with respect to subvectors $\boldsymbol{b}_i^r$. In case that $\boldsymbol{b}_i^r$ are common to all (except the last) categorical outcome values $k = 1, \ldots, K^r - 1$, the first derivative $\partial \ell \left(\boldsymbol{b}_i | \cdots \right) / \partial \boldsymbol{b}_i$ takes the following block form:

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

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

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

where $\cdots$ in expression above stands for corresponding elements of $-\boldsymbol{\Sigma}^{-(g)} \boldsymbol{b}_i$ coming from the prior distribution.

With regard to the Hessian matrix it is again better to deal with the two contributions separately. The basis of the negative Hessian matrix is formed by $\boldsymbol{\Sigma}^{-(g)}$. The other contribution comes in the form of a block-diagonal matrix, where the diagonal structure comes from the fact that $\boldsymbol{b}_i^r$ among different

outcomes $r \in \mathcal{R}$ do not interact within the model specification, i.e.,

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

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

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

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

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

B.4 Transformed ordered intercepts $\boldsymbol{a}$ for ordinal outcomes

The prior distribution for parameter $\boldsymbol{c}$ is specified through the probabilities $\boldsymbol{\pi}$ by (7). Specifying the prior for the probabilities allows for a more straightforward inclusion of prior knowledge. Both $\boldsymbol{c}$ and $\boldsymbol{\pi}$ cannot directly be used in combination with a Metropolis proposal without taking into account the limitation of the corresponding parametric space. Hence, we transform the parameter $\boldsymbol{a}$ in the following way:

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

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

$$c_k = \log \frac{e^{a_1} + \cdots + e^{a_k}}{e^{a_{k+1}} + \cdots + e^{a_{K-1}} + 1}, \quad a_k = \log \frac{\mathsf{logit}^{-1}(c_k) - \mathsf{logit}^{-1}(c_{k-1})}{1 - \mathsf{logit}^{-1}(c_{K-1})}$$

for $k = 1, \ldots, K-1$. Note that we dropped the outcome index r and the superscript (g) for the g-th cluster for simplicity. The prior (8) over $\boldsymbol{\pi}$ translates to the following form in terms of $\boldsymbol{a}$:

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

The logarithm of this can be easily differentiated:

$$\begin{aligned}
\log p(\boldsymbol{a}) &= \sum_{k=1}^{K} \alpha_k a_k - (\alpha_1 + \cdots + \alpha_K) \log \left(1 + \sum_{k=1}^{K-1} \exp\{a_k\}\right), \\
\frac{\partial \log p(\boldsymbol{a})}{\partial \boldsymbol{a}} &= \boldsymbol{\alpha} - (\alpha_1 + \cdots + \alpha_K)\, \mathsf{softmax}(\boldsymbol{a}), \\
-\frac{\partial^2 \log p(\boldsymbol{a})}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\top} &= (\alpha_1 + \cdots + \alpha_K) \\
&\qquad \left(\mathsf{diag}\{\mathsf{softmax}(\boldsymbol{a})\} - \mathsf{softmax}(\boldsymbol{a})\, \mathsf{softmax}(\boldsymbol{a})^\top\right).
\end{aligned}$$

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

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

Focusing on outcome $r \in \mathcal{R}^{\mathsf{Ord}}$ and group g we strip away the nuisance indices to obtain the expression

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References

Agresti A (2013) Categorical Data Analysis, 3rd edn. Wiley Series in Probability and Statistics, Wiley-Interscience

Brooks S, Gelman A, Jones G, et al (2011) Handbook for Markov Chain Monte Carlo, 2nd edn. Taylor & Francis, https://doi.org/10.1201/b10905

Fieuws S, Verbeke G (2004) Joint modelling of multivariate longitudinal profiles: Pitfalls of the random-effects approach. Statistics in Medicine 23:3093–3104. https://doi.org/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. https://doi.org/10.1111/j.1541-0420.2006.00507.x

Fitzmaurice G, Davidian M, Verbeke G, et al (2008) Longitudinal Data Analysis. CRC Press

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

Frühwirth-Schnatter S (2011) Dealing with Label Switching under Model Uncertainty, John Wiley & Sons, chap 10, pp 213–239

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

Hartzel J, Agresti A, Caffo B (2001) Multinomial logit random effects models. Statistical Modelling 1:81–102. https://doi.org/10.1177/1471082x0100100201

Komárek A, Komárková L (2013) Clustering for multivariate continuous and discrete longitudinal data. The Annals of Applied Statistics 7(1):177–200. https://doi.org/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. https://doi.org/10.18637/jss.v059.i12

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38(4):963–974. https://doi.org/10.2307/2529876

Malsiner-Walli G, Frühwirth-Schnatter S, Grün B (2016) Model-based clustering based on sparse finite Gaussian mixtures. Statistics and Computing 26:303–324. https://doi.org/10.1007/s11222-014-9500-2

Pinheiro JC, Chao EC (2006) Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics 15(1):58–81. https://doi.org/10.1198/106186006x96962

R Core Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org

Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26(2):195–239

Roeder K, Wasserman L (1997) Practical bayesian density estimation using mixtures of normals. Journal of the American Statistical Association 92(439):894–902

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

Therneau TM, Grambsch PM (2000) Modeling Survival Data: Extending the Cox Model. Springer-Verlag, New York

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

Wedel M, DeSarbo WS (1995) A mixture likelihood approach for generalized linear models. Journal of Classification 12(1):21–55