Journal of Classification
https://doi.org/10.1007/s00357-024-09466-2

Inferential Tools for Assessing Dependence Across Response Categories in Multinomial Models with Discrete Random Effects

Chiara Masci[1] · Francesca Ieva[1,2] · Anna Maria Paganoni[1]

Accepted: 12 February 2024

Abstract
We propose a discrete random effects multinomial regression model to deal with estimation and inference issues in the case of categorical and hierarchical data. Random effects are assumed to follow a discrete distribution with an a priori unknown number of support points. For a K-categories response, the modelling identifies a latent structure at the highest level of grouping, where groups are clustered into subpopulations. This model does not assume the independence across random effects relative to different response categories, and this provides an improvement from the multinomial semi-parametric multilevel model previously proposed in the literature. Since the category-specific random effects arise from the same subjects, the independence assumption is seldom verified in real data. To evaluate the improvements provided by the proposed model, we reproduce simulation and case studies of the literature, highlighting the strength of the method in properly modelling the real data structure and the advantages that taking into account the data dependence structure offers.

Keywords Discrete random effects · Multinomial regression · Unsupervised clustering · Multivariate statistics · Higher education

1 Introduction and Literature

This paper contributes to the existing literature on multinomial models, hierarchical data and unsupervised clustering by proposing an innovative multinomial model with discrete random effects.

In the framework of generalized linear models, multinomial outcomes have traditionally been treated separately from other response distributions. Indeed, it is more appropriate to consider multinomial models as multivariate generalized linear models, as they involve multiple logits for a multicategory response (Tutz & Hennevogl, 1996). This approach also applies

✉ Chiara Masci
chiara.masci@polimi.it

[1] MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, 20133 Milan, Italy

[2] Health Data Science Center, Human Technopole, Milan, Italy

Published online: 04 March 2024

when dealing with hierarchical data, such as longitudinal data or repeated measurements, where observations are naturally nested within groups (Agresti, 2018). Hierarchical data are commonly analyzed using linear mixed-effects models (Pinheiro & Bates, 2006), that incorporate in the linear predictor both *fixed effects*, associated to the entire population, and *random effects*, associated to the groups in which observations are nested, randomly drawn from the Gaussian-distributed population (Goldstein, 2011). Generalized linear mixed-effects models (GLMMs) deal with responses that follow distributions in the exponential family, other than Gaussian (Diggle et al., 2002; Agresti, 2018). However, extending GLMMs to handle unordered categorical responses presents more challenges (Daniels & Gatsonis, 1997; Hartzel & Agresti, 2001; Hedeker, 2003; Kuss et al., 2007; Wang & Tsodikov, 2010), due to the increased complexity associated with their modelling. An appropriate link function for nominal responses is the baseline-category logit, where fixed and random effects vary according to the response category.

As an alternative to the classical framework characterized by Gaussian-distributed random effects, we propose an innovative approach focused on the random effects of the multinomial model to provide new perspectives on their interpretation. We introduce a joint multinomial semi-parametric mixed-effects (JMSPME) model in which the random effects are assumed to follow a discrete distribution with an a priori unknown number of support points. For a multinomial response assuming K different categories, we assume the random effects to follow a joint $(K-1)$-variate discrete distribution where each of the $(K-1)$ marginal distributions can have a different number of support points. We adopt a baseline-category logit approach where fixed and random effects parameters are specific to each category.

This approach builds upon the recent literature on mixed-effects linear models with discrete random effects (Aitkin, 1999; Masci et al., 2019), which allow for the identification of a classification of highest-level units clustered into subpopulations based on their similarity in effects. Semi-parametric mixed-effects linear models (SLMMs) were initially proposed for continuous and binary responses (Masci et al., 2019, 2021; Maggioni, 2020), and a recent attempt has been made for multinomial responses (Masci et al., 2022), addressing a classification problem with hierarchical data related to engineering students at Politecnico di Milano (PoliMI). The authors in Masci et al. (2022) propose a multinomial semi-parametric model with discrete random effects with the aim to profile engineering students of Politecnico di Milano (PoliMI) into three categories (early dropout, late dropout and graduated), given some student personal and career information and considering their nested structure within engineering degree programmes. As the authors state, the main drawback of their method is the assumption of independent random effects relative to different response categories. This assumption simplifies the parameters estimation procedure, but it is a strong and seldom verified assumption, since the random effects of different logits arise from the same subjects. Our work mainly comes to tackle this issue as the proposed semi-parametric multinomial mixed-effects model does not assume the independence across the category-specific random effects distributions and is able to model different dependence structures across the multinomial categories. When estimating the parameters, this refinement results in two main advantages: the former is that we avoid bias in the estimates, induced by the natural dependence across categories; the latter is that, by jointly estimating the highest level units effects on the $K-1$ logits, we better investigate and interpret their trends. The drawback is that modelling the dependence across categories increases the dimensional complexity of the estimation procedure, requiring a nontrivial computational improvement.

The discrete random effects approach is also closely connected to a related field of study known as latent class and trait analysis (LCTA; Heinen, 1996). LCTA models aim to estimate latent traits, but there are key distinctions between our proposal and these models. Firstly, LCTA models require a predetermined number of latent classes, whereas our approach does not. Moreover, our approach handles covariates in a distinct manner compared to LCTA models.

We employ an estimation procedure based on an expectation-maximization (EM) algorithm inspired by the EM algorithm proposed in (Masci et al., 2022), relative to MSPEM. The model is also inserted in a clear inferential framework, that was lacking in previous works within this context. In particular, we complete the estimation procedure by adding the computation of the standard errors of the estimates and the assessment of the significance of the coefficients. The variance of maximum likelihood estimators is calculated by the inverse of the Fisher information matrix. For what concerns random effects significance, the variance partition coefficient (VPC) for a semi-parametric multinomial mixed-effects model is proposed.

In summary, the proposed model has several strengths (Rights & Sterba, 2016). Firstly, by assuming discrete random effects, we can identify a latent structure at the highest level of the hierarchy, which is a valuable alternative to the ranking provided by assuming Gaussian random effects. Secondly, the semi-parametric approach is more flexible and does not assume any specific parametric distribution, allowing for potential estimation of the true distribution of random effects. Thirdly, when the number of groups is extremely large, the identification of subpopulations can aid in interpreting the results and reducing dimensionality. Fourthly, identifying subpopulations provides insights into outlier detection, where the most populated subpopulations reveal the most frequent trends, while the smallest subpopulations contain groups with observations exhibiting anomalous behaviors compared to the majority. Lastly, from a practical perspective, assuming a discrete distribution for random effects allows us to express the likelihood as a weighted sum instead of a multiple integral, significantly simplifying the estimation procedure of the model parameters. Indeed, for a multinomial response, the integration issues typical of GLMMs to estimate the response marginal distribution become more complex (De Leeuw et al., 2008), requiring numerical approximations to evaluate the integral over the random effects distribution. The most commonly used methods are based on first- or second-order Taylor expansions (Goldstein & Rasbash, 1996), on a combination of a fully multivariate Taylor expansion and a Laplace approximation (Raudenbush et al., 2000), or using Gauss-Hermite quadrature (Stroud & Secrest, 1966; Hedeker, 2003). Nonetheless, these cited procedures are computationally very complex (McCulloch & Searle, 2001), and many authors have reported biased estimates using some of them (Breslow & Lin, 1995; Raudenbush et al., 2000; Rodríguez & Goldman, 1995). Specific software have been developed to perform these kinds of estimates—among others, HLM (Raudenbush, 2004), MLwiN (Steele et al., 2005), WinBugs (Spiegelhalter et al., 2003) and Supermix (Hedeker et al., 2008)—but they tend not to be very flexible and often require a high level of expertise of the users. Also, procedures implemented in SAS (PROC GLIMMIX, PROC NLMIXED; Caliński and Harabasz, 2013; Cary, 2015; Kuss et al., 2007), Stata (GSEM; Baum, 2016; Kuss et al., 2007) and SPSS (Corp., 2021) fit this type of models. In one of the most recent works on this topic (Hadfield, 2010), the authors develop a Markov chain Monte Carlo (MCMC) method for multi-response generalized linear mixed models, to provide a robust strategy for marginalizing the random effects (Zhao et al., 2006). This model is developed in a Bayesian setting—where the distinction between fixed and random effects does not technically exist—and the user should define the prior distributions on the parameters. The relative MCMCglmm R package (Hadfield, 2010) is, to the best of our knowledge, the only R package (R Core Team, 2019) that performs parametric mixed-effects multinomial regression.

To properly evaluate the performance of our model and its advantages with respect to its counterparts, we reproduce both simulation and case studies proposed in Masci et al. (2022), with a special focus on the comparison between JMSPME, MSPEM and the parametric MCMCglmm. Results show that JMSPME performs significantly better than MSPEM, and estimates are more accurate and have a reduced variance. Moreover, the subpopulations identified at the highest level of the hierarchy are much more coherent with the ranking estimated by the parametric MCMCglmm. These evidences frame the JMSPME model as the enhanced version of its antecedent MSPEM.

The remaining part of the paper is organized as follows: in Section 2, we describe the JMSPME model and its estimation procedure; in Section 3, we retrace the simulation study proposed in Masci et al. (2022) comparing the results of JMSPME and MSPEM; in Section 4, we apply the JMSPME model to the Politecnico di Milano case study presented in Masci et al. (2022), and we compare its results with the ones obtained by MSPEM and MCMCglmm; Section 5 draws the conclusions.

Software in the form of R code (R Core Team, 2021) together with the simulation study input data set and complete documentation is available on https://github.com/chiaramasci9/JMSPME.

2 Methodology: Joint Semi-parametric Mixed-Effects Model for a Multinomial Response

In this section, we first recall the basics of a mixed-effects multinomial model with discrete random effects (Section 2.1), and then, we present the JMSPME model and its algorithm (Section 2.2).

2.1 State of the Art: Multinomial Models with Discrete Random Effects

Consider a multinomial logistic regression model for nested data with a two-level hierarchy (Agresti, 2018; De Leeuw et al., 2008), where each observation j, for $j = 1, \ldots, n_i$, is nested within a group i, for $i = 1, \ldots, I$. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ be the n_i-dimensional response vector for observations within the $i-$th group. The multinomial distribution with K categories relative to Y_{ij} is the following:

$$Y_{ij} = \begin{cases} 1 & \pi_{ij1} \\ 2 & \pi_{ij2} \\ \ldots & \\ K & \pi_{ijK} \end{cases}, \tag{1}$$

where $k = 1, \ldots, K$ indexes the K support points of the discrete distribution of Y_{ij}, and π_{ijk} is the probability that observation j within group i assumes value k. Mixed-effects multinomial models assume that the probability that $Y_{ij} = k$, i.e., π_{ijk}, is given by

$$\begin{cases} \pi_{ijk} = P(Y_{ij} = k) = \dfrac{\exp(\eta_{ijk})}{1 + \sum_{k=2}^{K} \exp(\eta_{ijk})} & \text{for} \quad k = 2, \ldots, K \\ \pi_{ij1} = P(Y_{ij} = 1) = \frac{1}{1+\sum_{k=2}^{K} \exp(\eta_{ijk})} & \end{cases}, \tag{2}$$

where $\eta_{ijk} = \mathbf{x}_{ij}'\boldsymbol{\alpha}_k + \mathbf{z}_{ij}'\boldsymbol{\delta}_{ik}$ is the linear predictor. $\mathbf{x}_{ij}$ is the $p \times 1$ covariates vector (includes a 1 for the intercept) for the fixed effects, $\boldsymbol{\alpha}_k$ is the $p \times 1$ vector of fixed effects, $\mathbf{z}_{ij}$ is the $q \times 1$ covariates vector for the random effects (includes a 1 for the intercept) and $\boldsymbol{\delta}_{ik}$ is the $q \times 1$ vector of random effects. Logit models for nominal response basically pair each category with a baseline category. This formulation considers $K-1$ contrasts, between each category k, for $k = 2, \ldots, K$, and the reference category,[1] that is $k = 1$. Consequently, each category is assumed to be related to a latent "response tendency" for that category with respect to the reference one. Each contrast k', $k' = 1, \ldots, K-1$ is characterized by the set of contrast-specific parameters ($\boldsymbol{\alpha}_{k'}; \boldsymbol{\delta}_{ik'}$, for $i = 1, \ldots, I$), that models the probability of Y_{ij} being equal to $k \equiv k'+1$ with respect to the probability of Y_{ij} being equal to 1 (the reference category).[2] Starting from Eq. 2, the log-odds of each response with respect to the reference category are as follows:

$$\log\left(\frac{\pi_{ijk}}{\pi_{ij1}}\right) = \eta_{ijk} \qquad k = 2, \ldots, K. \tag{3}$$

For each contrast, the contrast-specific random effects describe the latent structure at the highest level of the hierarchy.

The maximum likelihood estimation (MLE) method allows to estimate the model parameters of this probability distribution.

Considering $\mathbf{A} = (\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K)$ and $\boldsymbol{\Delta}_i = (\boldsymbol{\delta}_{i2}, \ldots, \boldsymbol{\delta}_{iK})$, the distribution of Y_{ij}, conditional on the random effects distribution, takes the following form:

$$\begin{aligned} p(Y_{ij}|\mathbf{A}, \boldsymbol{\Delta}_i) =& \pi_{ij1}^{\mathbf{1}_{\{Y_{ij}=1\}}} \times \pi_{ij2}^{\mathbf{1}_{\{Y_{ij}=2\}}} \times \ldots \times \pi_{ijK}^{\mathbf{1}_{\{Y_{ij}=K\}}} = \\ =& \prod_{k=1}^{K} \pi_{ijk}^{\mathbf{1}_{\{Y_{ij}=k\}}} = \\ =& \prod_{k=1}^{K} \left(\frac{\exp(\eta_{ijk})}{1 + \sum_{l=2}^{K} \exp(\eta_{ijl})} \right)^{\mathbf{1}_{\{Y_{ij}=k\}}}. \end{aligned} \tag{4}$$

Assuming that Y_{ij} and $Y_{ij'}$ are independent for $j \neq j'$, the conditional distribution of $\mathbf{Y}_i$ is as follows:

$$\begin{aligned} p(\mathbf{Y}_i|\mathbf{A}, \boldsymbol{\Delta}_i) =& \frac{\left(\sum_{k=1}^{K}\left(\sum_{j=1}^{n_i} \mathbf{1}_{\{Y_{ij}=k\}}\right)\right)!}{\prod_{k=1}^{K}\left(\left(\sum_{j=1}^{n_i} \mathbf{1}_{\{Y_{ij}=k\}}\right)!\right)} \times \prod_{j=1}^{n_i} p(Y_{ij}|\mathbf{A}, \boldsymbol{\Delta}_i) = \\ =& \frac{\left(\sum_{k=1}^{K}\left(\sum_{j=1}^{n_i} \mathbf{1}_{\{Y_{ij}=k\}}\right)\right)!}{\prod_{k=1}^{K}\left(\left(\sum_{j=1}^{n_i} \mathbf{1}_{\{Y_{ij}=k\}}\right)!\right)} \times \prod_{j=1}^{n_i} \prod_{k=1}^{K} \pi_{ijk}^{\mathbf{1}_{\{Y_{ij}=k\}}} = \\ =& \frac{\left(\sum_{k=1}^{K}\left(\sum_{j=1}^{n_i} \mathbf{1}_{\{Y_{ij}=k\}}\right)\right)!}{\prod_{k=1}^{K}\left(\left(\sum_{j=1}^{n_i} \mathbf{1}_{\{Y_{ij}=k\}}\right)!\right)} \times \prod_{j=1}^{n_i} \prod_{k=1}^{K} \left(\frac{\exp(\eta_{ijk})}{1 + \sum_{l=2}^{K} \exp(\eta_{ijl})} \right)^{\mathbf{1}_{\{Y_{ij}=k\}}}. \end{aligned} \tag{5}$$

In the semi-parametric approach presented in Masci et al. (2022), the random effects are assumed to follow a discrete distribution with an a priori unknown number of support points

[1] We consider the first category as the reference one (here, $\eta_{ij1} = 1$), but this choice is arbitrary, and it does not affect the model formulation.

[2] Note that $k' \equiv k-1$ for $k = 2, \ldots, K$, and therefore, the sequence of parameters ($\boldsymbol{\alpha}_{k'}; \boldsymbol{\delta}_{ik'}$, for $i = 1, \ldots, I$) for $k' = 1, \ldots, K-1$ is equal to the sequence ($\boldsymbol{\alpha}_k; \boldsymbol{\delta}_{ik}$, for $i = 1, \ldots, I$) for $k = 2, \ldots, K$.

(Masci et al., 2019). Under this assumption, the multinomial logit takes the following form:

$$\eta_{ijk} = \mathbf{x}'_{ij}\boldsymbol{\alpha}_k + \sum_{m_k=1}^{M_k} \left[\mathbb{1}_{\{i \in m_k\}} \times \mathbf{z}'_{ij}\mathbf{b}_{m_k k} \right] \qquad m_k = 1, \dots, M_k, \quad k = 2, \dots, K, \tag{6}$$

where M_k is the total number of support points of the discrete distribution of $\mathbf{b}$ relative to the $k-$th category, for $k = 2, \dots, K$, and $\mathbb{1}\{i \in m_k\}$ indicates whether group i belongs to support point m_k. The random effects distribution relative to each category k, for $k = 2, \dots, K$, can be expressed as a set of points $(\mathbf{b}_{1k}, \dots, \mathbf{b}_{M_k k})$, where $M_k \leq I$ and $\mathbf{b}_{m_k k} \in \mathbb{R}^q$ for $m_k = 1, \dots, M_k$, and a set of weights $(w_{1k}, \dots, w_{M_k k})$, where $\sum_{m_k=1}^{M_k} w_{m_k k} = 1$ and $w_{m_k} \geq 0$:

$$\mathbf{B} = \begin{cases} \begin{cases} \mathbf{b}_{12}, & \mathbf{b}_{22}, \dots, \quad \mathbf{b}_{M_2 2} \\ (w_{12}), (w_{22}), \dots, (w_{M_2 2}) \end{cases} \\ \dots \\ \dots \\ \begin{cases} \mathbf{b}_{1K}, & \mathbf{b}_{2K}, \dots, \quad \dots, \quad \mathbf{b}_{M_K K} \\ (w_{1K}), (w_{2K}), \dots, \quad \dots, (w_{M_K K}) \end{cases} \end{cases} . \tag{7}$$

The discrete distributions P_k^*, for $k = 2, \dots, K$, belong to the class of all probability measures on $\mathbb{R}^q$ and are assumed to be independent. P_k^* is a discrete measure with M_k support points that can then be interpreted as the mixing distribution that generates the density of the stochastic model in Eq. 6. In particular, $w_{m_k k} = \mathrm{P}(\boldsymbol{\delta}_{ik} = \mathbf{b}_{m_k k})$, for $i = 1, \dots, I$. The maximum likelihood estimator $\hat{P}_k^*$ of P_k^* can be obtained following the theory of mixture likelihoods in Lindsay (1983); Lindsay et al. (1983), who proved the existence, discreteness and uniqueness of the semi-parametric maximum likelihood estimator of a mixing distribution, in the case of exponential family densities.

Given this formulation, Masci et al. (2022) propose an algorithm for implementing MSPEM for the joint estimations of $\boldsymbol{\alpha}_k$, $(\mathbf{b}_{1k}, \dots, \mathbf{b}_{M_k k})$ and $(w_{1k}, \dots, w_{M_k k})$, for $k = 2, \dots, K$, which is performed through the maximization of the complete likelihood. In the MSPEM's algorithm steps, under the independence assumption across the contrast-specific random effects, when estimating the support points relative to each contrast, the other contrast-specific random effects are fixed to the mean of the relative discrete distributions. In other words, when estimating the random effects of a group with respect to a response category, the random effects of this specific group with respect to the other categories are ignored. This assumption simplifies the parameter estimation procedure, but, as previously discussed, it is often too strict and unrealistic.

2.2 JMSPME Model

In the proposed JMSPME, we start from the modelling proposed in Eqs. 6 and 7, but we do not assume the independence across the random effects distributions relative to the $(K - 1)$ categories. Instead of considering $K - 1$ independent univariate discrete distributions, we refer to the $(K - 1)-$variate distribution of random effects. The object $\mathbf{B}$ defined in Eq. 7 is identified by a discrete distribution $\mathbf{P}^*$, that belongs to the class of all probability measures on $\mathbb{R}^{q \times (K-1)}$. $\mathbf{P}^*$ is a discrete measure with M_{tot} support points, where $M_{\text{tot}} = \prod_{k=2}^{K} M_k$ is the number of all possible combinations of the k-specific random effects $\mathbf{b}_{m_k k}$, for $m_k = 1, \dots, M_k$ and $k = 2, \dots, K$. We use $m = 1, \dots, M_{\text{tot}}$ to index the M_{tot} $(K - 1)-$variate support points and relative weights. By marginalizing this multivariate distribution, we then extract the marginal random effects distribution relative to each contrast k', for $k' = 1, \dots, K-1$.

The marginal likelihood is obtained as a weighted sum of all the conditional probabilities. In particular, the marginal likelihood of $\mathbf{Y}_i$ is the weighted sum of the likelihood of $\mathbf{Y}_i$ conditioned to all the M_{tot} possible combinations of the values of the $(K-1)$ discrete distributions of random effects:

$$L(\mathbf{Y}_i|\boldsymbol{A}) = \sum_{m=1}^{M_{\text{tot}}} w_m p(\mathbf{Y}_i|\boldsymbol{A}, \mathbf{B}_m). \tag{8}$$

w_m is the weight of the $m-$th combination of the $(K-1)$ weights distributions, and analogously, $\mathbf{B}_m$ is the $m-$th combination of the $(K-1)$ random effects distributions.

Under these assumptions, the JMSPME parameter estimates can be obtained by maximizing the likelihood in Eq. 8. Thanks to the likelihood convexity property, the maximization can be computed in two separate steps: one for computing the weights of the multivariate discrete distribution of the random effects and one for computing fixed effects and random effects iteratively.

The EM algorithm for the maximization of the two functionals is an iterative algorithm that alternates two steps (Dempster et al., 1977): the expectation step in which we compute the conditional expectation of the likelihood function with respect to the random effects, given the observations and the parameters that are computed in the previous iteration, and the maximization step in which we maximize the conditional expectation of the likelihood function. The observations are the values of the response variable y_{ij} and of the covariates $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$, for $j = 1, \dots, n_i$ and $i = 1, \dots, I$. The algorithm allows the number n_i, for $i = 1, \dots, I$, of observations to be different across groups, but within each group, missing data are not handled. The EM algorithm stops when the convergence or a maximum number of iterations are reached. As proven in Appendix A, the update for the parameters is given by the following:

$$w_m^{(\text{up})} = \frac{1}{I}\sum_{i=1}^{I} W_{im} \qquad m = 1, \dots, M_{\text{tot}}, \tag{9}$$

and

$$(\boldsymbol{A}^{(\text{up})}, \mathbf{b}_1^{(\text{up})}, \dots, \mathbf{b}_{(K-1)}^{(\text{up})}) = \arg\max_{\boldsymbol{A},\mathbf{B}_m} \sum_{m=1}^{M_{\text{tot}}} \sum_{i=1}^{I} W_{im} \times \ln p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{B}_m), \tag{10}$$

where each element w_m represents the weight of the $m-$th $(K-1)-$variate support point ($\mathbf{w}$ is an array with $K-1$ dimensions, i.e., a $(M_2 \times M_3 \times \dots \times M_K)-$ dimensional array) and, equivalently, W_{im} represents the probability that group i belongs to the $m-$th $(K-1)-$variate subpopulation, identified by the relative $K-1$ marginal subpopulations, conditionally on observations $\mathbf{y}_i$ and given the fixed effects $\boldsymbol{A}$ (W is an array with K dimensions, i.e., a $(I \times M_2 \times M_3 \times \dots \times M_K)-$ dimensional array of conditional weights[3]). In particular,

$$w_m = P(\boldsymbol{\Delta}_i = \mathbf{B}_m) \tag{11}$$

[3] Note that we are using a single index m to index a position in multidimensional objects (arrays w and W). We make this choice to ease the notation, calling with m the $m-$th combination of $(K-1)$ indices.

and

$$
\begin{aligned}
W_{im} =& \frac{w_m p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{B}_m)}{\sum_{\gamma=1}^{M_{\text{tot}}} w_\gamma p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{B}_\gamma)} \\
=& \frac{p(\boldsymbol{\Delta}_i = \mathbf{B}_m) p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{B}_m)}{p(\mathbf{y}_i|\boldsymbol{A})} \\
=& \frac{p(\mathbf{y}_i, \boldsymbol{\Delta}_i = \mathbf{B}_m|\boldsymbol{A})}{p(\mathbf{y}_i|\boldsymbol{A})} \\
=& p(\boldsymbol{\Delta}_i = \mathbf{B}_m|\mathbf{y}_i, \boldsymbol{A}) \qquad m = 1, \ldots, M_{\text{tot}}.
\end{aligned} \tag{12}
$$

By marginalizing W with respect to k, we obtain the marginal conditional weight matrices W_k, for $k = 2, \ldots, K$. The weight $w_m^{(\text{up})}$ is the mean of the conditional weights of the I groups, relative to the $m-$th $(K-1)-$variate support point.

The maximization step in Eq. 10 involves two steps, and it is done iteratively. In the first step, thanks to the convexity of the logarithm, for each category k, for $k = 2, \ldots, K$, we compute the arg max with respect to the support points $\mathbf{b}_{m_k k}$, for $m_k = 1, \ldots, M_k$, keeping $\boldsymbol{A}$ and $\mathbf{b}_l$, for $l \neq k$, fixed to the values computed at the previous iteration. In this way, we can maximize the expected log-likelihood (computed in the expectation step) with respect to all support points $\mathbf{b}_{m_k k}$ separately, i.e.,

$$
\begin{aligned}
\mathbf{b}_{m_k k}^{(\text{up})} =& \arg\max_{\mathbf{b}_k} \sum_{c_{m_k}=1}^{C_{\text{tot},k}} \sum_{i=1}^{I} W_{ic_{m_k}} \ln p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{b}_k, \mathbf{B}_{c_{m_k}}^{(\text{old})}) \\
& m_k = 1, \ldots, M_k, \quad k = 2, \ldots, K.
\end{aligned} \tag{13}
$$

where $C_{\text{tot},k} = M_{\text{tot}}/M_k$ is the number of $(K-1)-$variate support points that have m_k as marginal support point for category k. $W_{ic_{m_k}}$ represents the probability that group i belongs to the latent subpopulation c_{m_k}, that is identified by m_k, relatively to category k, and the support points relative to the other $K-2$ categories that correspond to the $c_{m_k}-$th combination. $\mathbf{B}_{c_{m_k}}^{(\text{old})}$ is the set of random effects, estimated at the previous iteration, relative to categories $(2, \ldots, k-1, k+1, \ldots, K)$, that compose the $c_{m_k}-$th combination with m_k. In particular,

$$
p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{b}_k, \mathbf{B}_{c_{m_k}}^{(\text{old})}) = \frac{\left(\sum_{\gamma=1}^{K}\left(\sum_{j=1}^{n_i} \mathbf{1}_{\{y_{ij}=\gamma\}}\right)\right)!}{\prod_{\gamma=1}^{K}\left(\left(\sum_{j=1}^{n_i} \mathbf{1}_{\{y_{ij}=\gamma\}}\right)!\right)} \times \prod_{j=1}^{n_i} \prod_{\gamma=1}^{K} \left(\frac{\exp(\eta_{ij\gamma})}{1+\sum_{\nu=2}^{K} \exp(\eta_{ij\nu})}\right)^{\{\mathbf{1}_{y_{ij}=\gamma}\}}, \tag{14}
$$

where

$$
\eta_{ij\gamma} = \begin{cases} \mathbf{x}_{ij}'\boldsymbol{\alpha}_k + \mathbf{z}_{ij}'\mathbf{b}_k & \text{if} \quad \gamma = k \\ \mathbf{x}_{ij}'\boldsymbol{\alpha}_\gamma + \mathbf{z}_{ij}'\mathbf{b}_{(m_\gamma \gamma)_{c_{m_k}}}^{(\text{old})} & \text{if} \quad \gamma \neq k \end{cases}. \tag{15}
$$

$\mathbf{b}_{(m_\gamma \gamma)_{c_{m_k}}}^{(\text{old})}$ are the random effects relative to the support point $(m_\gamma \gamma)_{c_{m_k}}$, that is the support point relative to category γ that compose the $c_{m_k}-$th combination with m_k.

In the second step, we fix the support points of the random effects distributions computed in the previous step, and we compute the arg max in Eq. 10 with respect to $\mathbf{A}$. Again, thanks to the convexity of the logarithm, we can compute the arg max in Eq. 10 with respect to $\boldsymbol{\alpha}_k$, separately for each $k = 2, \ldots, K$, keeping $\boldsymbol{\alpha}_l$, for $l \neq k$ fixed to the values computed at the previous iteration.

To compute the point $\mathbf{B}_m$ for each group i, for $i = 1, \ldots, I$, we maximize the conditional probability of $\boldsymbol{\Delta}_i$ given the observations $\mathbf{y}_i$ and the fixed effects $\boldsymbol{A}$. The estimates of random effects $\mathbf{\Delta}_i$ for each group i, $i = 1, \dots, N$, are obtained by maximizing W_{im} over m, i.e.,

$$\hat{\mathbf{\Delta}}_i = \mathbf{B}_{\tilde{m}} \qquad \text{where} \quad \tilde{m} = \arg\max_m W_{im}$$

MSPEM's and JMSPME's algorithm skeletons are very similar. Nonetheless, substantial differences regard the estimation of the random effects distribution, i.e., of the weights (Eq. 9) and of the random effects support points (Eq. 13). In the MSPEM's algorithm, only marginal weights and marginal conditional weight matrices are treated, and in the maximization in Eq. 13, the latent subpopulation structure relative to the other categories is ignored. In the JMSPME's algorithm, all weights and conditional weights are treated in their multivariate setting, and the function to be maximized in Eq. 13 takes into account the conditional weights of groups across all categories. The multivariate optimization implies an increased computational cost, that scales with the number of covariates and of response categories.

During the iterations of the EM algorithm, the reduction of the support points of the random effects discrete distributions is performed. In particular, if two support points are closer, in terms of Euclidean distance, than a pre-specified tuning parameter D_k, for $k = 1, \dots, K$, they collapse to a unique point. The algorithm converges when the difference between the estimates of fixed and random effects obtained at two consecutive iterations is smaller than fixed tolerance thresholds (`tollF` and `tollR`, respectively). The initialization of the support points of the discrete distribution and the choice of the tuning parameter D_k, for $k = 1, \dots, K$, are two key aspects of the EM algorithm, and a detailed description about their tuning can be found in Appendix B. Further details about convergence criteria and model identifiability can be found in Masci et al. (2019) and Masci et al. (2022).

Besides the point estimates of both fixed and random effects, a further improvement provided by the algorithm for the *JMSPME* model regards the computation of their standard errors and the assessment of their significance. The variance of maximum likelihood estimators is calculated by the inverse of the Fisher information matrix. Consider $\boldsymbol{\theta}$ the parameter vector to be estimated and $\hat{\boldsymbol{\theta}}_{ML}$ its ML estimate:

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\theta}}_{ML}) &= [I(\hat{\boldsymbol{\theta}}_{ML})^{-1}] \\ &= (-\text{E}[H(\hat{\boldsymbol{\theta}}_{ML})])^{-1} \\ &= \left(-\left[\frac{\partial^2 ln\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ML}}\right)^{-1}.\end{aligned}$$

$H(\hat{\boldsymbol{\theta}}_{ML})$ is the Hessian matrix, i.e., the matrix of second derivatives of the likelihood $\mathcal{L}$ with respect to the parameter $\boldsymbol{\theta}$, evaluated in $\hat{\boldsymbol{\theta}}_{ML}$. The second derivatives of the observed-data log-likelihood are obtained by numerical differentiation (Meng et al., 1991). The standard error of each estimator is just the square root of this estimated variance (King, 1989; Long et al., 1997). For what concerns random effects variance, given the estimated support points $b_{m_k k}$, for $m_k = 1, \dots, M_k$ and $k = \{2, 3\}$, and relative weights, the variance σ^2_{rk} of the two marginal distributions of random effects can be computed, thanks to the law of total variances, as

$$\sigma^2_{rk} = \text{Var}[B_k] = \text{E'}[\text{Var}[B_k|(b_{1k}, \dots, b_{M_k k})]] + \text{Var}[\text{E'}[B_k|(b_{1k}, \dots, b_{M_k k})]], \qquad (16)$$

where

$$\begin{aligned}
&\mathrm{E}[\mathrm{Var}[B_k|(b_{1k},\dots,b_{M_k k})]] = \\
&\qquad = \mathrm{E}[B_k^2|(b_{1k},\dots,b_{M_k k})] - (\mathrm{E}[B_k|(b_{1k},\dots,b_{M_k k})])^2 = \\
&\qquad\qquad = \sum_{m_k=1}^{M_k} b_{m_k k}^2 \times w_{m_k k} - \left(\sum_{m_k=1}^{M_k} b_{m_k k} \times w_{m_k k}\right)^2
\end{aligned}$$

and, assuming $b_{m_k k}$, for $m_k = 1,\dots,M_k$, to be independent

$$\mathrm{Var}[\mathrm{E}[B_k|(b_{1k},\dots,b_{M_k k})]] = \mathrm{Var}\left[\sum_{m_k=1}^{M_k} b_{m_k k} \times w_{m_k k}\right] = \sum_{m_k=1}^{M_k} \mathrm{Var}[b_{m_k k}] \times w_{m_k k}^2.$$

The covariance between the couple B_k and B_h, for each $h, k = 2,\dots,K$, can be computed as follows:

$$\begin{aligned}
\mathrm{Cov}(B_h, B_k) &= \mathrm{E}[B_h B_k] - \mathrm{E}[B_h]\mathrm{E}[B_k] = \\
&= \sum_{m=1}^{M_h \times M_k} w_m \times b_{mh} \times b_{mk} - \left(\sum_{m_h=1}^{M_h} w_{m_h h} \times b_{m_h h}\right) \times \left(\sum_{m_k=1}^{M_k} w_{m_k k} \times b_{m_k k}\right).
\end{aligned} \tag{17}$$

3 Simulation Study

In this section, we reproduce the simulation study proposed in Masci et al. (2022), and we compare the performances of JMSPME and MSPEM models. A categorical response variable assuming $K = 3$ possible values is considered, where $k = 1$ is the reference category. Three different settings are simulated: (i) one considering only a random intercept, (ii) one considering only a random slope and (iii) one considering both random intercept and random slope.[4]

$I = 100$ groups of data are considered, where each group contains 200 observations.[5] Data are simulated in order to induce the presence of three subpopulations regarding category $k = 2$, i.e., $M_2 = 3$, and two subpopulations regarding category $k = 3$, i.e., $M_3 = 2$. In particular, for $j = 1,\dots,200$ and $i = 1,\dots,100$, the model is

$$\begin{aligned}
\pi_{ijk} &= P(Y_{ij} = k) = \frac{\exp(\eta_{ijk})}{1 + \sum_{l=2}^{3} \exp(\eta_{ijl})} \qquad \text{for} \quad k = 2, 3; \\
\pi_{ij1} &= P(Y_{ij} = 1) = \frac{1}{1 + \sum_{l=2}^{3} \exp(\eta_{ijl})},
\end{aligned} \tag{18}$$

[4] Masci et al. (2022) make this choice since in the application for modelling student dropout, the model considers a three-categories response and only a random intercept. In the simulation study, they maintain the three-categories response, to ease the reader, and, besides the case (i) of a random intercept, they add the other two random effects cases, in order to show the model results in more complex settings. They also include two covariates in the model (considered both for fixed effects or one for random and one for fixed) to be in line with the case study.

[5] The number of observations is allowed to be different across groups. Here, to facilitate the reader and without loss of generality, they are taken equally across groups.

where the linear predictor $\boldsymbol{\eta}_{ik} = (\eta_{i1k}, \ldots, \eta_{i200k})$ is generated according to the following data generating process (DGP)[6]:

(i) Random intercept case ($\boldsymbol{\eta}_{ik} = \alpha_{1k}\mathbf{x}_{1i} + \alpha_{2k}\mathbf{x}_{2i} + \delta_{ik}$)

$$\boldsymbol{\eta}_{i2} = \begin{cases} +4\mathbf{x}_{1i} - 3\mathbf{x}_{2i} - 7 & i = 1, \ldots, 30 \\ +4\mathbf{x}_{1i} - 3\mathbf{x}_{2i} - 4 & i = 31, \ldots, 60 \\ +4\mathbf{x}_{1i} - 3\mathbf{x}_{2i} - 2 & i = 61, \ldots, 100 \end{cases}$$

$$\boldsymbol{\eta}_{i3} = \begin{cases} -2\mathbf{x}_{1i} + 2\mathbf{x}_{2i} - 5 & i = 1, \ldots, 60 \\ -2\mathbf{x}_{1i} + 2\mathbf{x}_{2i} - 2 & i = 61, \ldots, 100 \end{cases} \tag{19}$$

(ii) Random slope case ($\boldsymbol{\eta}_{ik} = \alpha_{1k} + \alpha_{2k}\mathbf{x}_{1i} + \delta_{ik}\mathbf{z}_{1i}$)

$$\boldsymbol{\eta}_{i2} = \begin{cases} -1 - 3\mathbf{x}_{1i} + 5\mathbf{z}_{1i} & i = 1, \ldots, 30 \\ -1 - 3\mathbf{x}_{1i} + 2\mathbf{z}_{1i} & i = 31, \ldots, 60 \\ -1 - 3\mathbf{x}_{1i} - 1\mathbf{z}_{1i} & i = 61, \ldots, 100 \end{cases}$$

$$\boldsymbol{\eta}_{i3} = \begin{cases} -2 + 2\mathbf{x}_{1i} - 2\mathbf{z}_{1i} & i = 1, \ldots, 60 \\ -2 + 2\mathbf{x}_{1i} - 6\mathbf{z}_{1i} & i = 61, \ldots, 100 \end{cases} \tag{20}$$

(iii) Random intercept and slope case ($\boldsymbol{\eta}_{ik} = \alpha_k\mathbf{x}_{1i} + \delta_{1ik} + \delta_{2ik}\mathbf{z}_{1i}$)

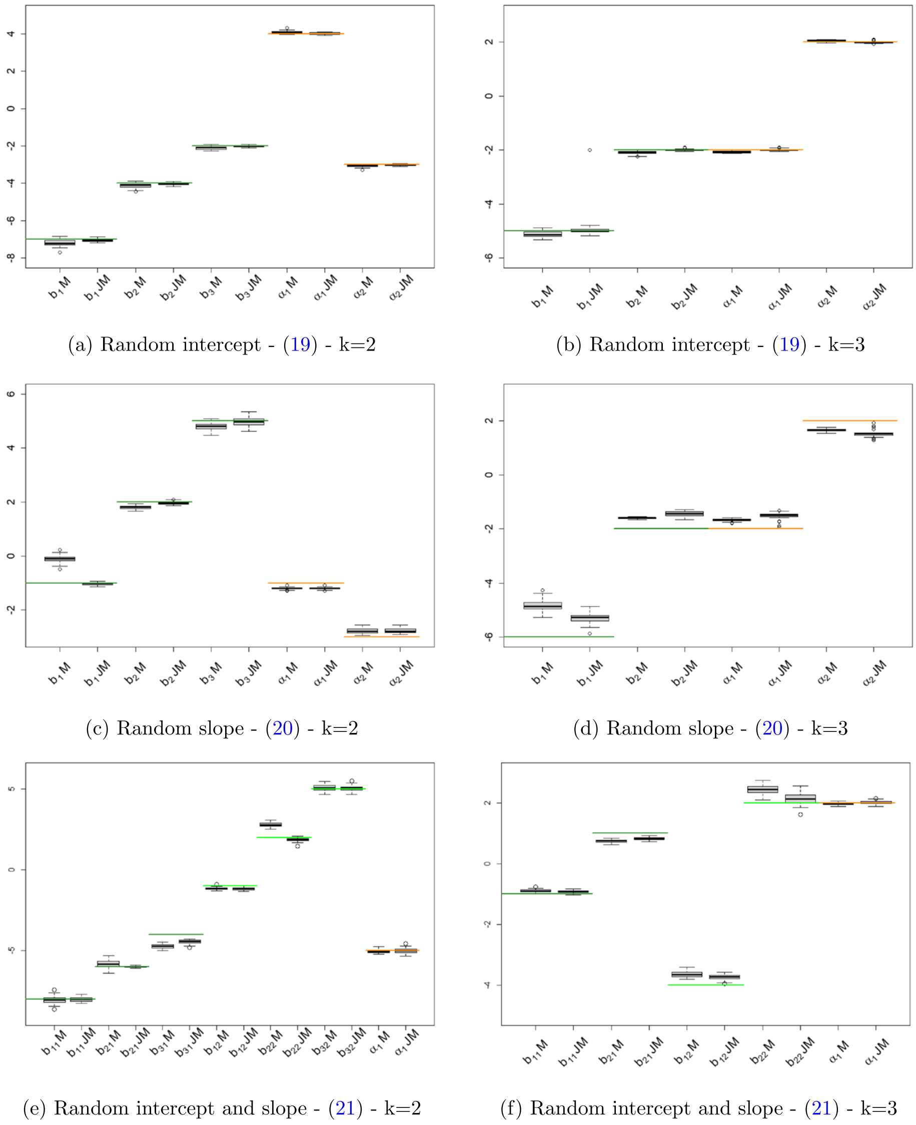
$$\boldsymbol{\eta}_{i2} = \begin{cases} -5\mathbf{x}_{1i} - 6 + 5\mathbf{z}_{1i} & i = 1, \ldots, 30 \\ -5\mathbf{x}_{1i} - 4 + 2\mathbf{z}_{1i} & i = 31, \ldots, 60 \\ -5\mathbf{x}_{1i} - 8 - 1\mathbf{z}_{1i} & i = 61, \ldots, 100 \end{cases}$$

$$\boldsymbol{\eta}_{i3} = \begin{cases} +2\mathbf{x}_{1i} + 1 - 4\mathbf{z}_{1i} & i = 1, \ldots, 60 \\ +2\mathbf{x}_{1i} - 1 + 2\mathbf{z}_{1i} & i = 61, \ldots, 100 \end{cases} \tag{21}$$

Variables $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{z}_1$ are normally distributed with mean equal to 0 and standard deviation equal to 1.

All the parameters used to simulate the data and the tuning parameters of the semi-parametric model are equal to the ones in Masci et al. (2022). In particular, we perform 500 runs of the JMSPME model for each of the three settings shown in Eqs. 19, 20 and 21. We fix $D_k = 1$ as a threshold value for the support point collapse criterion, for $k = \{2, 3\}$ and $\texttt{tollR} = \texttt{tollF} = 0.01$ for the convergence (see Appendix B in Masci et al. (2022) for the details). In all the runs, the algorithm of the JMSPME model converges in a number of iterations that ranges between 4 and 7, slightly quicker with respect to MSPEM, whose number of iterations ranges between 5 and 10. Nonetheless, the computational time for one run of JMSPME's algorithm is higher than the one for MSPEM. Indeed, in this simulation study, JMSPME's algorithm takes on average 60 min for a run relative to DGP (i) and 100 min for a run relative to DGPs (ii) and (iii). For the same runs, MSPEM takes about 30 and 60 min. Figure 1 reports the distribution of JMSPME and MSPEM estimated fixed and random effects, across the runs, for the three DGPs. Descriptive statistics of the estimates are also reported in Tables 6 and 7 in Appendix C.

[6] Without loss of generality, we consider two covariates, simulating the case in which they are both for fixed effects or one for random and one for fixed. The choice of fixed and random effects values is arbitrary: in this case, they are chosen in order to simulate different situations in which we obtain both balanced and unbalanced categories.



(a) Random intercept - (19) - k=2

(b) Random intercept - (19) - k=3

(c) Random slope - (20) - k=2

(d) Random slope - (20) - k=3

(e) Random intercept and slope - (21) - k=2

(f) Random intercept and slope - (21) - k=3

Fig. 1 Distribution of MSPEM and JMSPME estimates in the simulation study, across runs, for the three DGPs. Green and orange lines mark the real random and fixed effects, respectively. MSPEM and JMSPME estimates are identified with “M”and “JM”, respectively

Fixed effects estimates are evaluated on the total number of runs, while random effects ones are evaluated on the runs in which the estimated number of subpopulations corresponds to the simulated one (that is the majority of the cases). Note that, when the algorithm identifies a higher number of subpopulations with respect to the simulated ones, it simply splits a subpopulation into two or more subpopulations, but the fixed effects estimates do not result to be affected by the number of subpopulations identified in the data. Estimates result to be very accurate, both for fixed and random effects, and their variability across runs is substantially low. Compared to MSPEM, the JMSPME model produces more precise and stable estimates.

We observe a 93.55%, 64.12% and 51.43% decrease in the mean estimation error, for the three settings, respectively. Moreover, given that the ML estimates in multinomial regression are only *asymptotically* unbiased, we expect the performance of the algorithm to increase when the number of observations increases (Masci et al., 2022).

For what concerns the identification of the subpopulations, Table 1 reports the number of runs out of 500 in which the two models identify the simulated number of subpopulations (i.e., $M_2 = 3$ and $M_3 = 2$) and correctly assign groups to the identified subpopulations, for all the three DGPs.

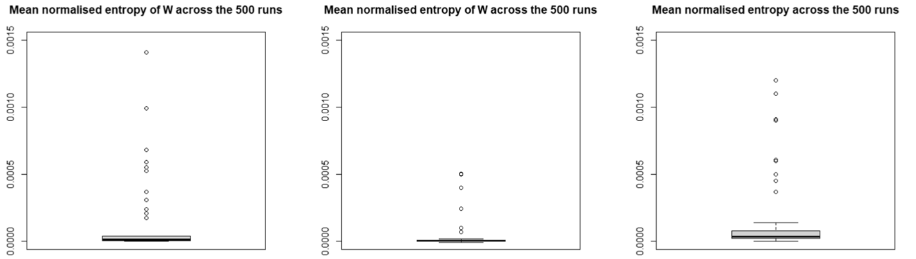
Except for the case (ii), the JMSPME's algorithm correctly identifies the simulated number of subpopulations and classifies groups into these subpopulations in a higher number of runs with respect to the MSPEM's algorithm. In the random slope case, the two models identify the correct number of subpopulations with approximately the same incidence, but the JMSPME's algorithm shows a better performance in assigning groups to the identified subpopulations.

Extending the approach presented in Masci et al. (2022) into our multivariate setting, we evaluate the uncertainty of the classification of groups into subpopulations by measuring, for each group, the normalized entropy of the conditional weight distribution. By looking at the three-dimensional array W, we evaluate the uncertainty of classification of each group into one of the M_{tot} $(K-1)-$variate subpopulations. In order to compute the response category-specific uncertainty of classification, in the MSPEM approach, the authors consider the marginal conditional weight matrices W_k, for $k = \{2, \dots, K\}$. Here, we compute the global uncertainty of the classification of each group, with respect to all response categories, by looking at the K-dimensional array W. The normalized entropy of each first-dimension i of the array W is computed as the entropy $E_i = -\sum_{m=1}^{M_{\text{tot}}} W_{im} \log(W_{im})$ divided by the maximum possible entropy value relative to M_{tot} subpopulations, i.e., $-\log(1/M_{\text{tot}})$. We recall that the lowest level of uncertainty is reached when the algorithm assigns a group to a bivariate subpopulation m, with probability 1; in this case, the normalized entropy of the group would be equal to 0. On the opposite, the highest level of uncertainty is reached when the distribution of the conditional weights of a group i is uniform on the M_{tot} subpopulations ($W_{im} = 1/M_{\text{tot}}$ for $m{=}1, \dots, M_{\text{tot}}$), which corresponds to an entropy $E_i = -\log(1/M_{\text{tot}})$, and, therefore, to a normalized entropy equal to 1. The normalized entropy constitutes also a driver for the choice of the tuning parameters D_k (details in Masci et al. (2022)). Figure 2 reports the distribution of the normalized entropy of W_i, for $i = 1, \dots, I$, for the three simulated cases, mediated on the runs in which the JMSPME's algorithm identifies $M_2 = 3$ and $M_3 = 2$.

Table 1 JMSPME and MSPEM models performances across 500 runs for each of the three DGPs

	Number of runs in which the model identifies $M_2 = 3$ and $M_3 = 2$		Number of runs in which the model correctly classifies all groups into subpopulations	
	MSPEM	JMSPME	MSPEM	JMSPME
(i) Random intercept case	473 out of 500	480 out of 500	470 out of 473	471 out of 480
(ii) Random slope case	453 out of 500	452 out of 500	427 out of 453	452 out of 452
(iii) Random intercept and slope case	422 out of 500	460 out of 500	315 out of 422	400 out of 460

The first two columns report the number of runs out of 500 in which the algorithms identify the correct number of subpopulations that are simulated in DGPs in Eqs. 19, 20 and 21; third and fourth columns report the number of runs, out of the number of runs in which the algorithms identify $M_2 = 3$ and $M_3 = 2$, in which the algorithms correctly assign each group to the correspondent subopulation



(a) Random intercept case (b) Random slope case (c) Random intercept & slope case

Fig. 2 Boxplots of the normalized entropy of W, measured for each group, obtained by mediating the entropy in the runs in which the algorithm identifies $M_2 = 3$ and $M_3 = 2$, for the random intercept case (**a**), random slope case (**b**) and random intercept and slope case (**c**)

We observe that the entropy level is always very low (considering that the maximum normalized entropy is 1), suggesting that, for the simulated data, the JMSPME's algorithm clearly distinguishes the presence of patterns within the data. The normalized entropy computed on the runs in which the algorithm identifies a higher number of subpopulations is, as expected, higher: since the algorithm estimates two very close subpopulations instead of the single simulated one, it does not clearly distinguish the belonging of groups into these subpopulations.

4 Case Study: University Student Dropout Across Engineering Degree Programmes

The main novelty introduced by the JMSPME is twofold. The former regards the ability to take into account and model the correlation structure across response category-specific random effects; the latter regards the positioning of the model in a tailored inferential framework.

In order to test and evaluate these aspects in the real data application that drove the development of this type of models, we reproduce the case study presented in Masci et al. (2022), and we compare our results with the ones obtained by both the MSPEM and the parametric MCMCglmm models.

4.1 Data and Model Setting

The case study consists in the application of the model to data about Politecnico di Milano (PoliMI) students, in order to classify different profiles of engineering students and to identify subpopulations of similar degree programmes. The sample includes the concluded careers of students enrolled in some of the engineering programmes of PoliMI in the academic year between 2010/2011 and 2015/2016. The dataset considers 18, 604 concluded careers of students nested within 19 engineering degree programmes (the smallest and the largest degree programmes contain 341 and 1246 students, respectively). 32.7% of these careers is concluded with a dropout, while the remaining 67.3% regards graduate students. The response variable regards the status of the concluded career that can be classified as follows:

- *Graduate*: occurs when the student concludes his/her career obtaining the bachelor's degree (67.3% of the sample)

- *Early dropout*: occurs when the student drops within the third semester after the enrolment (16.2% of the sample)
- *Late dropout*: occurs when the student drops after the third semester after the enrolment (16.5% of the sample)

The distinction between the two types of dropout is motivated by the interest in distinguishing the determinants that drive them, that might be structurally different and approached by different mitigation strategies.

Regarding student characteristics, besides the status of the concluded career and the degree programme the student is enrolled in, the number of European credit transfer system credits (ECTS), i.e., the credits he/she obtained at the first semester of the first year of career (the variable has been standardized in order to have 0 mean and 1 sd) and his/her gender (the sample contains 22.3% females and 77.7% males), is considered.[7] Table 2 reports the variables considered in the analysis with their explanation.[8]

The modelling proposed is the following. For each student j, for $j = 1, \ldots, n_i$, nested within degree programme i, for $i = 1, \ldots, I$ (with $I = 19$), the mixed-effects multinomial logit model takes the following form:

$$Y_{ij} = \begin{cases} \text{Graduate} & \pi_{ij1} \\ \text{Early dropout} & \pi_{ij2} \\ \text{Late dropout} & \pi_{ij3} \end{cases}, \tag{22}$$

where

$$\pi_{ijk} = P(Y_{ij} = k) = \frac{\exp(\eta_{ijk})}{1 + \sum_{k=2}^{3} \exp(\eta_{ijk})} \qquad \text{for} \quad k = 1, \ldots, 3 \tag{23}$$

and

$$\eta_{ijk} = \begin{cases} \mathbf{x}'_{ij}\boldsymbol{\alpha}_k + \delta_{ik} & k = 2, 3 \\ 0 & k = 1 \end{cases}. \tag{24}$$

Y_{ij} corresponds to the student `Status` (graduate is the reference category); $\mathbf{x}_{ij}$ is the two-dimensional vector of covariates for the fixed effects, that contains student `Gender` and `TotalCredits1.1`; $\boldsymbol{\alpha}_k$ is the two-dimensional vector of fixed effects relative to the k-th category; and δ_{ik} is the random intercept relative to the i-th degree programme (`DegProg`) and to the k-th category.

Given the data setting and model formulation presented in Eqs. 22, 23, 24, we apply the three models to PoliMI data. The aim of the study is to model the probability of being an early or late dropout student, with respect to being a graduate one, given student characteristics and early career information, and considering the nested structure of students within the 19-degree programmes. Both MSPEM's and JMSPME's algorithms, by assuming discrete random effects, identify subpopulations of degree programmes, depending on their effects on early and late dropout probability, while the MCMCglmm's algorithm, by assuming Gaussian random effects, identifies a ranking of degree programmes.

The MSPEM algorithm assumes the two effects of each degree programme on early and late dropout probability to be independent, while in the JMSPME's algorithm, we assume there is an unknown dependence structure.

[7] Masci et al. (2022) state that only information at the first semester of career is used because it is observable for all students (either dropout or graduate) and it allows to predict student dropout as soon as possible, i.e., at the beginning of the student career.

[8] For further information on the original dataset and the preprocessing phase, please refer to Masci et al. (2022).

Table 2 List and explanation of variables at the student level to be included in the model (Masci et al., 2022)

Variable	Description	Type of variable
`Status`	Status of concluded career	Three-level factor (G = graduate; $D1$ = early dropout; $D2$ = late dropout)
`Gender`	Gender of the student	Binary (male=0, female=1)
`TotalCredits1.1`	Number of ECTS obtained by the student during the first semester of the first year	Continuous
`DegProg`	Degree programme the student is enrolled in	19-level factor

4.2 JMSPME, MSPEM and MCMCglmm Results

We run JMSPME and MSPEM models' algorithms with the same parameter setting chosen in Masci et al. (2022): `tollR=tollF=`10^{-2}, `itmax`=60, `it1`=20, $\tilde{w} = 0$ and $D_k = 0.3$, for $k = 2, 3$. JMSPME's algorithm converges in 9 iterations (computational time $\approx$ 21 min) and identifies 5 supopulations for both categories $k = 2$ (early dropout) and $k = 3$ (late dropout). MSPEM's algorithm converges in 7 iterations (computational time $\approx$ 13 min) and identifies 4 supopulations for both categories.

Fixed Effects Estimates

For what concerns fixed effects, Table 3 reports the estimates obtained for the three models.

MSPEM and JMSPME estimated parameters are very close to each other and coherent with the MCMCglmm ones. The estimated significant fixed effects of JMSPME and MCMCglmm coincide.[9] In particular, both JMSPME and MCMCglmm results show that females have, on average, a lower probability of late dropout with respect to males, while no significant gender difference emerges for early dropout, and that the number of credits obtained at the first semester is inversely proportional to the probability of both early and late dropout. Fixed effects result to be robust and invariant with respect to different random effect assumptions.

Random Effects Estimates

Regarding our main interest, the analysis of random effects, the estimated discrete distributions of JMSPME and MSPEM are reported in Table 4 and displayed in Fig. 3. Figure 4 reports the degree programmes ranking estimated by MCMCglmm, complemented by the correspondent JMSPME and MSPEM mass points. The list of degree programmes belonging to the various subpopulations, for both JMSPME and MSPEM, can be obtained from Fig. 4 and is explicitly provided in Tables 8 and 9, in Appendix D.

Discrete random effects are reported in increasing order and are numbered accordingly. For JMSPME, each subpopulation is identified by one out of five possible values of the intercept relative to $k = 2$ and one out of five possible values of the intercept relative to $k = 3$. JMSPME identifies biomedical engineering as the degree programme in which students are more likely to early drop, all else equal, while civil and environmental engineering and environmental and land planning engineering result to be the ones in which students tend to early drop less than

[9] MSPEM's algorithm does not include any measurement of standard errors or coefficients significance.

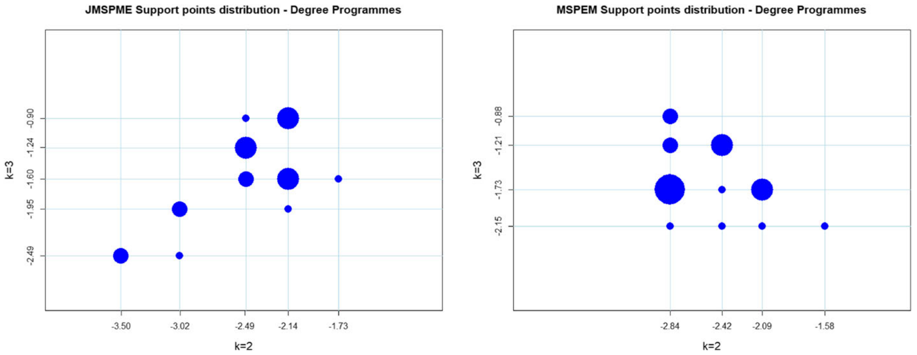
Table 3 Estimated fixed effects of JMSPME, MSPEM and MCMCglmm models for student dropout prediction

	k=2		k=3	
	$\hat{\alpha}_1$ (Gender)	$\hat{\alpha}_2$ (TotalCredits1.1)	$\hat{\alpha}_1$ (Gender)	$\hat{\alpha}_2$ (TotalCredits1.1)
JMSPME	0.014	-2.684^{***}	-0.577^{***}	-1.907^{***}
	(0.0609)	(0.0218)	(0.0606)	(0.0211)
MSPEM	-0.153	-2.704	-0.685	-1.899
MCMCglmm	-0.027	-2.797^{**}	-0.6234^{**}	-2.135^{**}
	(0.0403)	(0.0444)	(0.0862)	(0.0321)

For JMSPME and MCMCglmm, standard errors of the estimates are reported in brackets. Asterisks denote different levels of significance: $0.01 < p$-value < 0.1; $^*0.001 < p$-value < 0.01; $^{**}0.0001 < p$-value < 0.001; $^{***}p$-value < 0.0001

Table 4 Estimated random effects of JMSPME and MSPEM models for student dropout prediction. For JMSPME, standard errors of the estimates are reported in brackets

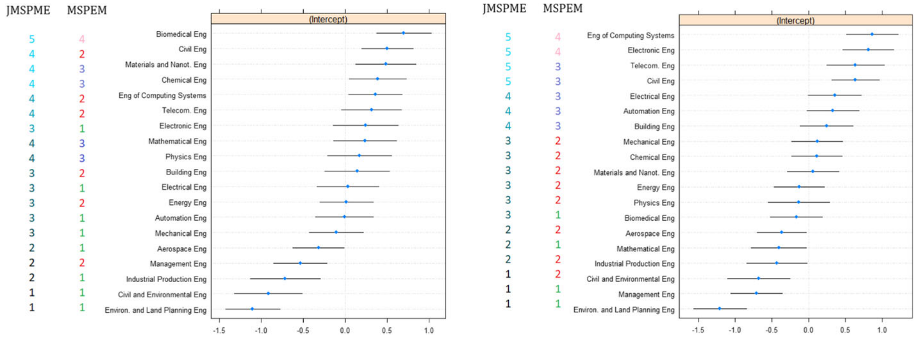
	k=2		k=3	
	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$
JMSPME	$\hat{b}_{12} = -3.504(0.0780)$	$\hat{w}_{12} = 0.100$	$\hat{b}_{13} = -2.491(0.1001)$	$\hat{w}_{13} = 0.147$
	$\hat{b}_{22} = -3.023(0.0710)$	$\hat{w}_{22} = 0.167$	$\hat{b}_{23} = -1.950(0.0566)$	$\hat{w}_{23} = 0.173$
	$\hat{b}_{32} = -2.485(0.0385)$	$\hat{w}_{32} = 0.291$	$\hat{b}_{33} = -1.601(0.0386)$	$\hat{w}_{33} = 0.321$
	$\hat{b}_{42} = -2.138(0.0537)$	$\hat{w}_{42} = 0.391$	$\hat{b}_{43} = -1.245(0.0519)$	$\hat{w}_{43} = 0.144$
	$\hat{b}_{52} = -1.728(0.0429)$	$\hat{w}_{52} = 0.051$	$\hat{b}_{53} = -0.903(0.0430)$	$\hat{w}_{53} = 0.215$
MSPEM	$\hat{b}_{12} = -2.841$	$\hat{w}_{12} = 0.482$	$\hat{b}_{13} = -2.152$	$\hat{w}_{13} = 0.210$
	$\hat{b}_{22} = -2.423$	$\hat{w}_{22} = 0.272$	$\hat{b}_{23} = -1.733$	$\hat{w}_{23} = 0.421$
	$\hat{b}_{32} = -2.096$	$\hat{w}_{32} = 0.193$	$\hat{b}_{33} = -1.219$	$\hat{w}_{33} = 0.262$
	$\hat{b}_{42} = -1.586$	$\hat{w}_{42} = 0.053$	$\hat{b}_{43} = -0.880$	$\hat{w}_{43} = 0.107$



(a) JMSPME degree programmes distribution across subpopulations.

(b) MSPEM degree programmes distribution across subpopulations.

Fig. 3 Estimated discrete bivariate random effects distributions of JMSPME (**a**) and MSPEM (**b**). Bubbles are centred in the random effects estimates, and their size is proportional to the number of degree programmes belonging to the subpopulations



(a) Early Dropout

(b) Late Dropout

Fig. 4 **a** and **b** The ranking of the MCMCglmm estimated intercepts with their confidence intervals relative to the 19-degree programmes, for early (k=2) and late (k=3) dropout, respectively. Alongside degree programme names, we report subpopulation indexes estimated by both JMSPME's and MSPEM's algorithms. Colours are only intended to help in the visualization

the others, all else equal (Table 8). These two subpopulations have relatively lower weights with respect to the other three subpopulations, that represent the majority of the sample, and, consequently, are interpreted as the ones containing three-degree programmes with anomalous behaviors. For late dropout, degree programmes are more uniformly distributed across the five subpopulations, starting from Subpopulation 1, that contains the three-degree programmes associated with the lowest late dropout probability, until Subpopulation 5, that contains the four-degree programmes associated with the highest late dropout probability.

Comparing the subpopulations identified by the two semi-parametric models, we notice some differences. Both MSPEM and JMSPME identify biomedical engineering as the degree programme in which students are more likely to early drop. For late dropout, both MSPEM and JMSPME assign civil and environmental engineering and environmental and land planning engineering to the subpopulation associated with the highest late dropout probability. The remaining of the distributions of degree programmes on the estimated subpopulations is more heterogeneous across the two models. What is interesting to note is the comparison between the distributions of degree programmes on the bivariate subpopulations of MSPEM and JMSPME, displayed in Fig. 3. Each bubble size is proportional to the weight of the bivariate subpopulation. For JMSPME, the distribution of the weights on the bisector of the figure in panel a suggests that, except for very few cases (e.g., biomedical engineering), degree programme effects are quite coherent between early and late dropout. On the opposite, the distribution of the weights of the bivariate subpopulations obtained by MSPEM (panel b) suggests that degree programmes in which students are more likely to early drop are less likely to late drop and vice versa. This result demonstrates that different assumptions on the dependence structure across random effects distributions lead to relevant differences in the estimates and in their interpretation.

Regarding the comparison with the parametric MCMCglmm approach, Fig. 4 shows that the subpopulations identified by JMSPME align closely with the rankings obtained from MCMCglmm. However, the results from MSPEM do not exhibit the same level of precision and coherence in matching the ranking with the subpopulations, as it only provides a partial alignment.

Goodness of Fit and *JMSPME* Evaluation

To compare the models in terms of goodness of fit, we compute their relative misclassification tables (Table 5).

Error rates are 22.1% for JMSPME, 21.6% for MCMCglmm and 23.3% for MSPEM, respectively. As noted in Masci et al. (2022), we expect the MCMCglmm to have the best fit, since it estimates a single random effect for each degree programme (and, therefore, it fits the data 'deeply'). JMSPME error rate is lower than the MSPEM one, and it is very close to the MCMCglmm one, suggesting that the identified subpopulations catch almost the entire heterogeneity across degree programme effects. This is somehow expected since the more flexible assumptions of JMSPME and MCMCglmm result in a better capacity to model the real dynamics within the data. Given the high predictive performance and the matching with the parametric approach, the JMSPME's algorithm proves to produce precise and reliable estimates and to overperform compared to MSPEM.

Finally, to complement the JMSPME results, we evaluate the uncertainty of classification and the model PVRE. The uncertainty of classification is evaluated by measuring, for each degree programme $i = 1, \ldots, 19$, the normalized entropy of the conditional weights, computed as $E_i = -\sum_{m=1}^{M_{\text{tot}}} W_{im} \log(W_{im})$ divided by the maximum possible entropy value relative to M_{tot} subpopulations, i.e., $-\log(1/M_{\text{tot}})$, where $M_{\text{tot}} = 5 \times 5 = 25$. The mean and median of the 19 normalized entropy distributions are 0.0785 and 0.0426, respectively, while minimum and maximum values are 0.0002 and 0.2681, respectively, indicating a low level of uncertainty of classification.

Besides the support points and relative weights of the two marginal discrete distributions of random effects B_2 and B_3 reported in Table 4, we estimate their variance-covariance matrix, the correlation between B_2 and B_3 and the VPCs. For $k = \{2, 3\}$, we compute $\sigma_{r2}^2 = 0.2275 + 0.0146 = 0.2421$ and $\sigma_{r3}^2 = 0.2603 + 0.0006 = 0.2609$, as defined in Eq. 16.

In order to compute the covariance, we refer to the estimated 5×5-matrix of joint weights $\mathbf{w}$[10]

$$\mathbf{w} = \begin{bmatrix} 0.0947 & 0.0049 & 0 & 0 & 0 \\ 0.0526 & 0.1130 & 0.0013 & 0 & 0 \\ 0 & 0.0044 & 0.1167 & 0.1355 & 0.0348 \\ 0 & 0.0511 & 0.1524 & 0.0081 & 0.1797 \\ 0 & 0 & 0.0508 & 0 & 0 \end{bmatrix}.$$

and, by following Eq. 17, we compute $\text{Cov}(B_2, B_3) = 4.1553 - (-2.5024) \times (-1.5916) = 0.1725$. The variance-covariance matrix of $\mathbf{B}$ is, therefore,

$$\text{Var}(B_2, B_3) = \begin{pmatrix} 0.2421 & 0.1725 \\ 0.1725 & 0.2609 \end{pmatrix}$$

and the correlation between B_1 and B_2 is 0.6863, that is in line with what we expected by looking at panel a in Fig. 3. Lastly, the VPC relative to each logit $k = \{2, 3\}$, that is the portion of the total variability in the response explained by the latent structure identified at the degree programme level, is evaluated as

$$\text{VPC}_k = \frac{\sigma_{rk}^2}{\sigma_{rk}^2 \times \pi^2/3} = \begin{cases} 0.06857 & \text{for} \quad k = 2 \\ 0.07348 & \text{for} \quad k = 3 \end{cases}.$$

[10] Rows and columns refer to the support points as ordered in Table 4, for $k = \{2, 3\}$, respectively.

Table 5 Misclassification tables relative to JMSPME (left tabular), MSPEM (central tabular) and MCMCglmm (right tabular) predictions, expressed in percentages

	obs D1	obs D2	obs G	obs D1	obs D2	obs G	obs D1	obs D2	obs G
pred D1	0.099	0.060	0.018	0.095	0.063	0.019	0.100	0.058	0.018
pred D2	0.033	0.043	0.014	0.035	0.038	0.017	0.032	0.047	0.017
pred G	0.032	0.064	0.637	0.032	0.066	0.635	0.030	0.061	0.637

For both early and late dropout, about 7% of the total variability is explained by the subpopulation structure. Results of MCMCglmm provide $\text{VPC}_2 = 0.0906$ and $\text{VPC}_3 = 0.1091$.

5 Concluding Remarks and Future Perspectives

In this paper, we propose an enhanced version of a mixed-effects model with discrete random effects for unordered multinomial responses, called JMSPME, together with a suitable inferential framework. Estimates of parameters are obtained through an EM algorithm. The JMSPME consists in a semi-parametric approach that assumes the response category-specific random effects to follow a discrete distribution with an a priori unknown number of mass points, that are allowed to differ across response categories. With respect to the traditional parametric approach, the JMSPME model constitutes a valid alternative, both from a computational and an interpretative point of view. Indeed, the discrete distribution on the random effects allows to write the likelihood function as a weighted sum, avoiding integration issues typical of parametric mixed-effects multinomial models, and, moreover, allows to identify a latent structure of subpopulations at the highest level of grouping.

JMSPME, in which we do not consider any independence assumption across response-specific random effects distributions, is an improvement on its previous MSPEM model presented in Masci et al. (2022). By relaxing this seldom verified assumption considered in MSPEM, JMSPME results to be a more powerful model, able to provide more accurate and less uncertain estimates and to better model the heterogeneity at the highest level of the grouping.

In order to test and evaluate the performances of the JMSPME model, compared to the MSPEM ones, we reproduce the simulation and case studies reported in Masci et al. (2022). JMSPME overperforms MSPEM both in the simulation and case studies and the introduction of the inferential framework results to be a further value added that adds interpretability.

In the context of predicting the types of concluded careers of PoliMi students, nested within different engineering degree programmes, the JMSPME model proves higher predictive performance compared to MSPEM, and the estimated subpopulations of degree programmes, that differ from the ones estimated by MSPEM, are extremely coherent with the ranking obtained by applying the parametric MCMCglmm.

All these evidences support the thesis that, in discrete random effects multinomial models, a joint modelling of the random effects distributions across response categories is paramount and overrides the previous version of the model.

This paper enters both in the literature about multinomial regression (Agresti, 2018) and in the one about mixed-effects models with discrete random effects (Aitkin, 1999; Hartzel & Agresti, 2001; Masci et al., 2019). The proposed model contributes to both the streams but, at the same time, suffers from some of their typical criticalities. Given the presence of multiple logits, multinomial regression models are treated as multivariate models, and in addition, the likelihood function is such that its maximization in closed form is not feasible. These two aspects contribute to require an important computing power and numerical methods for the maximization steps. For what concerns mixed-effects models with discrete random effects, we believe that they are extremely useful in many different contexts of application and that the research of a latent structure of subpopulations at the highest level of grouping is an innovative and informative way of analyzing this level of the hierarchy. Their application to real data in which the cardinality of the groups is very high and in which subpopulations are a posteriori explained can provide important insights. Nonetheless, although these methods do not require to fix the number of subpopulations a priori but they estimate it together with the other parameters, this estimate is extremely sensitive to the choice of the threshold distance D. Some criteria to choose D have been proposed in the literature (Masci et al., 2019, 2022), but its choice is still sensitive and influential. For these reasons, future work will be devoted to the embedding of more efficient optimization algorithms and to the development of a clear rule to drive the choice of the threshold distance D.

The JMSPME model can be applied to any classification problem dealing with an unordered categorical response and hierarchical data, a context in which the statistical literature is still poor and quite challenging. Its extension to deal with ordinal responses could be a further interesting development.

Appendix A: Proof of increasing likelihood property

In the EM algorithm proposed in Section 2.2, the updates of the parameters are obtained in order to increase the likelihood, such that:

$$L(\mathbf{A}^{(\mathrm{up})}|\mathbf{y}) \geq L(\mathbf{A}|\mathbf{y}),$$

where $\mathbf{A}^{(\mathrm{up})}$ are the updated fixed effects, and the likelihood $L(\mathbf{A}^{(\mathrm{up})}|\mathbf{y})$ is computed summing up the random effects with respect to the updated discrete distribution $(\mathbf{B}_m^{(\mathrm{up})}, w_m^{(\mathrm{up})})$ for $m = 1, \ldots, M_{\mathrm{tot}}$. Thanks to the definition of the likelihood function in Eq. 8, we have that:

$$\log\left(\frac{L(\boldsymbol{A}^{(\mathrm{up})}|\mathbf{y})}{L(\boldsymbol{A}|\mathbf{y})}\right) = \sum_{i=1}^{I} \log\left(\frac{p(\mathbf{y}_i|\boldsymbol{A}^{(\mathrm{up})})}{p(\mathbf{y}_i|\boldsymbol{A})}\right).$$

All these terms can be bounded below by the quantity:

$$\log\left(\frac{p(\mathbf{y}_i|\boldsymbol{A}^{(\mathrm{up})})}{p(\mathbf{y}_i|\boldsymbol{A})}\right) \geq Q_i(\theta^{(\mathrm{up})}, \theta) - Q_i(\theta, \theta), \tag{25}$$

where

$$Q_i(\theta^{(\mathrm{up})}, \theta) = \sum_{m=1}^{M_{\mathrm{tot}}} \left(\frac{w_m p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{B}_m)}{p(\mathbf{y}_i|\boldsymbol{A})}\right) \log(w_m^{(\mathrm{up})} p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{B}_m)).$$

$Q_i(\theta, \theta)$ is analogously defined and $\theta = (\boldsymbol{A}, \mathbf{B}_1, \ldots, \mathbf{B}_{M_{\mathrm{tot}}}, w_1, \ldots, w_M)$. This bound can be found thanks to the convexity of the logarithm (proof in Azzimonti et al. (2013)).

Defining

$$Q(\theta^{(\mathrm{up})}, \theta) = \sum_{i=1}^{I} Q_i(\theta^{(\mathrm{up})}, \theta) \quad \text{and} \quad Q(\theta, \theta) = \sum_{i=1}^{I} Q_i(\theta, \theta),$$

we obtain, thanks to Eq. 25, a lower bound for the quantity of interest

$$\log\left(\frac{L(\boldsymbol{A}^{(\mathrm{up})}|\mathbf{y})}{L(\boldsymbol{A}|\mathbf{y})}\right) \geq Q(\theta^{(\mathrm{up})}, \theta) - Q(\theta, \theta).$$

In order to show now that $\forall\theta$, $Q(\theta^{(\mathrm{up})}, \theta) \geq Q(\theta, \theta)$, we can show that, $\forall\theta$ fixed, $\theta^{(\mathrm{up})}$ is defined as the $\arg\max_{\tilde{\theta}} Q(\tilde{\theta}, \theta)$.

Defining W_{im} as the probability that the $i-$th group belongs to the $m-$th combination among the M_{tot} possible combinations, conditionally on the observations $\mathbf{y}_i$ and given the fixed effects parameters $\boldsymbol{A}$, we obtain

$$\begin{aligned}
Q(\tilde{\theta}, \theta) =& \sum_{i=1}^{I} \sum_{m=1}^{M_{\mathrm{tot}}} \left(\frac{w_m p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{B}_m)}{p(\mathbf{y}_i|\boldsymbol{A})}\right) \log(\tilde{w}_m p(\mathbf{y}_i|\tilde{\boldsymbol{A}}, \tilde{\mathbf{B}}_m)) \\
=& \sum_{i=1}^{I} \sum_{m=1}^{M_{\mathrm{tot}}} W_{im} \log(\tilde{w}_m p(\mathbf{y}_i|\tilde{\boldsymbol{A}}, \tilde{\mathbf{B}}_m)) \\
=& \sum_{i=1}^{I} \sum_{m=1}^{M_{\mathrm{tot}}} W_{im} \log(\tilde{w}_m) + \sum_{i=1}^{I} \sum_{m=1}^{M} W_{im} \log(p(\mathbf{y}_i|\tilde{\boldsymbol{A}}, \tilde{\mathbf{B}}_m)) \\
=& J_1(\tilde{w}_1, \ldots, \tilde{w}_{M_{\mathrm{tot}}}) + J_2(\tilde{\boldsymbol{A}}, \tilde{\mathbf{B}}_1, \ldots, \tilde{\mathbf{B}}_{M_{\mathrm{tot}}}).
\end{aligned} \tag{26}$$

The functionals J_1 and J_2 can be maximized separately. In particular, by maximizing the functional J_1, we obtain the updates for the weights of the random effects distribution, and by maximizing the functional J_2 in an iterative way, we obtain the estimates of $\boldsymbol{A}$ and $\mathbf{B}_m$, for $m = 1, \ldots, M_{\mathrm{tot}}$.

Appendix B: EM algorithm technical details

Two key aspects of the *JMSPME* model's algorithm regard the initialization of the support points of the discrete distribution and the choice of the tuning parameter D_k for the support reduction procedure. In this appendix, we report our initialization procedure, and we discuss possible choices for D_k.

Support Point Initialization

The EM algorithm is extremely sensitive to the initial grid on which we identify the random effects discrete distribution. For this reason, we follow an initialization procedure that aims to be inclusive and not binding. In particular, the algorithm starts by considering N support points for the random effects that are estimated by fitting N distinct multinomial regression models (one for each group i). The weights are uniformly distributed on these N support points. A valuable alternative is to fit a classical multilevel multinomial model, with data nested within the N groups, where both the intercept and the slope are treated as random effects. This approach allows the algorithm to start from a very capillary setting and to perform a tailored dimensional reduction. When N is extremely large, starting from N mass points could be time-consuming and not strictly necessary. In this situation, the user can estimate $N^* < N$ support points by computing the range of the parameters estimated by the N multinomial regression models (or, by the single multilevel multinomial model) and, then, randomly extracting $N^* < N$ points from a uniform distribution on the range. Again, the weights of the initial grid are uniformly distributed on the N^* points.

This procedure results to be robust and generalizable, and it allows to reach stable and good estimates in both the simulation and the case studies.

Tuning of the Parameter D_k

D_k is certainly the most sensitive parameter of the algorithm. As the value of D_k increases, the likelihood of mass points collapsing also increases, resulting in a decrease in the final number of support points. The choice of the parameter D_k might be driven by different factors:

- *A priori knowledge*: In situations where users are specifically seeking subpopulations with parameters that differ by a specific quantity that can be interpreted, D_k can be set accordingly. Although this is the simplest scenario, it is less commonly encountered.
- *Evaluation of the entropy*: As mentioned in the main text, the entropy of the conditional weights matrix W_k can serve as a useful indicator for selecting an appropriate value for D_k. It can be argued that good values of D_k are those that result in the algorithm classifying groups with low uncertainty, which corresponds to a low entropy of the matrix W_k.
- *Identification of stability regions*: In the absence of prior knowledge, a further potential approach to select the optimal value for D_k is to experiment with multiple values and assess the trend in the number of identified subpopulations. Ideally, D_k should encompass a range of values that result in the identification of a reasonably high number of subpopulations, gradually decreasing until values that lead to the identification of a single mass point. This iterative process is helpful to highlight stability regions, i.e., ranges of values of D_k for which the algorithm identifies the same number of masses. Similar to hierarchical clustering, stability regions identify the existence of distinct patterns within the data.
- *The elbow method*: The final and well-known method that can provide valuable insights into determining the optimal choice of D_k involves evaluating the model likelihood and other goodness-of-fit indices such as BIC (Bayesian information criterion) and AIC (Akaike information criterion). We anticipate that as the number of mass points decreases, the likelihood of the model will also decrease. However, following the concept of the elbow method, we are interested in identifying the point where the decrease in model likelihood is minimal for the lowest number of mass points. This approach helps identify the optimal value of D_k that balances model complexity and goodness of fit.

In addition to recommending the optimal value for D_k, these procedures offer additional valuable insights regarding the performance of the model, the robustness of the results, and the strength of the interpretation.

Appendix C: JMSPME and MSPEM results of the simulation study

Table 6 Estimated fixed and random effects of JMSPME model in the three different DGPs of the simulation study

	$\hat{\alpha}_{1k}$	$\hat{\alpha}_{2k}$	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_{12} = 4.002 \pm 0.085$	$\hat{\alpha}_{22} = -2.998 \pm 0.080$	$\hat{b}_{12} = -7.009 \pm 0.152$	$\hat{w}_{12} = 0.300$
			$\hat{b}_{22} = -4.006 \pm 0.084$	$\hat{w}_{22} = 0.300$
			$\hat{b}_{32} = -2.012 \pm 0.061$	$\hat{w}_{32} = 0.400$
	$TV = +4$	$TV = -3$	$TV = (-7, -4, -2)$	$TV = (0.3, 0.3, 0.4)$
k=3	$\hat{\alpha}_{13} = -1.994 \pm 0.038$	$\hat{\alpha}_{23} = 2.005 \pm 0.037$	$\hat{b}_{13} = -5.016 \pm 0.091$	$\hat{w}_{13} = 0.599$
			$\hat{b}_{23} = -2.004 \pm 0.048$	$\hat{w}_{23} = 0.401$
	$TV = -2$	$TV = +2$	$TV = (-5, -2)$	$TV = (0.6, 0.4)$

Estimated fixed and random effects of JMSPME model for the DGP in Eq. 19.

	$\hat{\alpha}_{1k}$	$\hat{\alpha}_{2k}$	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_{12} = -0.993 \pm 0.038$	$\hat{\alpha}_{22} = -2.963 \pm 0.079$	$\hat{b}_{12} = 4.964 \pm 0.143$	$\hat{w}_{12} = 0.300$
			$\hat{b}_{22} = 1.946 \pm 0.053$	$\hat{w}_{22} = 0.301$
			$\hat{b}_{32} = -1.017 \pm 0.052$	$\hat{w}_{32} = 0.399$
	$TV = -1$	$TV = -3$	$TV = (+5, +2, -1)$	$TV = (0.3, 0.3, 0.4)$
k=3	$\hat{\alpha}_{13} = -1.873 \pm 0.029$	$\hat{\alpha}_{23} = 1.859 \pm 0.049$	$\hat{b}_{13} = -1.699 \pm 0.156$	$\hat{w}_{13} = 0.600$
			$\hat{b}_{23} = -5.307 \pm 0.289$	$\hat{w}_{23} = 0.400$
	$TV = -2$	$TV = +2$	$TV = (-2, -6)$	$TV = (0.6, 0.4)$

Estimated fixed and random effects of JMSPME model for the DGP in Eq. 20.

	$\hat{\alpha}_{k}$	$\hat{b}_{1m_k k}$	$\hat{b}_{2m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_{2} = -5.007 \pm 0.125$	$\hat{b}_{112} = -5.982 \pm 0.057$	$\hat{b}_{212} = 5.032 \pm 0.146$	$\hat{w}_{12} = 0.300$
		$\hat{b}_{122} = -4.459 \pm 0.118$	$\hat{b}_{222} = 1.827 \pm 0.136$	$\hat{w}_{22} = 0.300$
		$\hat{b}_{132} = -8.011 \pm 0.129$	$\hat{b}_{232} - 1.147 \pm 0.097$	$\hat{w}_{32} = 0.400$
	$TV = -5$	$TV = (-6, -4, -8)$	$TV = (+5, +2, -1)$	$TV = (0.3, 0.3, 0.4)$
k=3	$\hat{\alpha}_{3} = 2.021 \pm 0.048$	$\hat{b}_{113} = 0.836 \pm 0.047$	$\hat{b}_{213} = -3.742 \pm 0.092$	$\hat{w}_{13} = 0.600$
		$\hat{b}_{123} = -0.917 \pm 0.044$	$\hat{b}_{223} = 2.139 \pm 0.165$	$\hat{w}_{23} = 0.400$
	$TV = +2$	$TV = (+1, 1)$	$TV = (-4, +2)$	$TV = (0.6, 0.4)$

Estimated fixed and random effects of JMSPME model for the DGP in Eq. 21.

Estimates are reported in terms of mean ± sd, computed on the 500 runs of the simulation study for the fixed effects and on the runs in which the algorithm identifies $M_2 = 3$ and $M_3 = 2$ (reported in Table 1) for the random effects. True values (TV) of the coefficients used to simulate data are reported under the relative estimates

Table 7 Fixed and random effects estimated by MSPEM in the three different DGPs of the simulation study, first presented in Masci et al. (2022) and reproduced in this paper

	$\hat{\alpha}_{1k}$	$\hat{\alpha}_{2k}$	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_{12} = 4.096 \pm 0.081$	$\hat{\alpha}_{22} = -3.051 \pm 0.053$	$\hat{b}_{12} = -6.819 \pm 0.182$	$\hat{w}_{12} = 0.300$
			$\hat{b}_{22} = -3.916 \pm 0.109$	$\hat{w}_{22} = 0.300$
			$\hat{b}_{32} = -2.122 \pm 0.099$	$\hat{w}_{32} = 0.400$
	$TV = +4$	$TV = -3$	$TV = (-7, -4, -2)$	$TV = (0.3, 0.3, 0.4)$
k=3	$\hat{\alpha}_{13} = -2.067 \pm 0.046$	$\hat{\alpha}_{23} = 2.059 \pm 0.034$	$\hat{b}_{13} = -5.200 \pm 0.089$	$\hat{w}_{13} = 0.599$
			$\hat{b}_{23} = -1.899 \pm 0.048$	$\hat{w}_{23} = 0.401$
	$TV = -2$	$TV = +2$	$TV = (-5, -2)$	$TV = (0.6, 0.4)$

Fixed and random effects estimated by MSPEM algorithm for the DGP in Eq. 19.

	$\hat{\alpha}_{1k}$	$\hat{\alpha}_{2k}$	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_{12} = -1.195 \pm 0.039$	$\hat{\alpha}_{22} = -2.766 \pm 0.085$	$\hat{b}_{12} = 4.786 \pm 0.121$	$\hat{w}_{12} = 0.300$
			$\hat{b}_{22} = 1.811 \pm 0.071$	$\hat{w}_{22} = 0.301$
			$\hat{b}_{32} = -0.117 \pm 0.134$	$\hat{w}_{32} = 0.399$
	$TV = -1$	$TV = -3$	$TV = (+5, +2, -1)$	$TV = (0.3, 0.3, 0.4)$
k=3	$\hat{\alpha}_{13} = -1.672 \pm 0.039$	$\hat{\alpha}_{23} = 1.713 \pm 0.051$	$\hat{b}_{13} = -1.601 \pm 0.057$	$\hat{w}_{13} = 0.600$
			$\hat{b}_{23} = -4.791 \pm 0.210$	$\hat{w}_{23} = 0.400$
	$TV = -2$	$TV = +2$	$TV = (-2, -6)$	$TV = (0.6, 0.4)$

Fixed and random effects estimated by MSPEM algorithm for the DGP in Eq. 20.

	$\hat{\alpha}_{k}$	$\hat{b}_{1m_k k}$	$\hat{b}_{2m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_{2} = -5.013 \pm 0.098$	$\hat{b}_{112} = -5.863 \pm 0.236$	$\hat{b}_{212} = 5.091 \pm 0.195$	$\hat{w}_{12} = 0.300$
		$\hat{b}_{122} = -4.700 \pm 0.129$	$\hat{b}_{222} = 2.801 \pm 0.119$	$\hat{w}_{22} = 0.300$
		$\hat{b}_{132} = -8.022 \pm 0.237$	$\hat{b}_{232} = -1.185 \pm 0.079$	$\hat{w}_{32} = 0.400$
	$TV = -5$	$TV = (-6, -4, -8)$	$TV = (+5, +2, -1)$	$TV = (0.3, 0.3, 0.4)$
k=3	$\hat{\alpha}_{3} = 1.977 \pm 0.040$	$\hat{b}_{113} = 0.739 \pm 0.058$	$\hat{b}_{213} = -3.651 \pm 0.092$	$\hat{w}_{13} = 0.600$
		$\hat{b}_{123} = -0.888 \pm 0.055$	$\hat{b}_{223} = 2.419 \pm 0.160$	$\hat{w}_{23} = 0.400$
	$TV = +2$	$TV = (+1, -1)$	$TV = (-4, +2)$	$TV = (0.6, 0.4)$

Fixed and random effects estimated by MSPEM algorithm for the DGP in Eq. 21.

Estimates are reported in terms of mean ± sd, computed on the 500 runs of the simulation study for the fixed effects and on the runs in which the algorithm identifies $M_2 = 3$ and $M_3 = 2$ (shown in Table 1 in Masci et al. (2022) and here reported in Table 1) for the random effects. True values (TV) of the coefficients used to simulate data are reported under the relative estimates

Appendix D: Distribution of the 19-degree programmes across the MSPEM and JMSPME identified subpopulations

Table 8 JMSPME: distribution of the 19-degree programmes across the five identified subpopulations relative to $k = 2, 3$

Subpopulation 1	Subpopulation 2	Subpopulation 3	Subpopulation 4	Subpopulation 5
Early dropout (*k*=2)				
Civil and environmental Eng	Aerospace Eng	Building Eng	Chemical Eng	Biomedical Eng
Environ. and Land Planning Eng	Industrial Production Eng	Automation Eng	Civil Eng	
	Management Eng	Electrical Eng	Materials and Nanot. Eng	
		Electronic Eng	Telecom. Eng	
		Energy Eng	Physics Eng	
		Mechanical Eng	Eng of Computing Systems	
			Mathematical Eng	
Late dropout (*k*=3)				
Civil and Environmental Eng	Aerospace Eng	Biomedical Eng	Building Eng	Civil Eng
Management Eng	Industrial Production Eng	Chemical Eng	Automation Eng	Telecom. Eng
Environ. and Land Planning Eng	Mathematical Eng	Materials and Nanot. Eng	Electrical Eng	Electronic Eng
		Energy Eng		Eng of Computing Systems
		Physics Eng		
		Mechanical Eng		

For each k, subpopulations are ordered from 1 to 5 coherently with the estimated coefficients reported in Table 4

Table 9 MSPEM: distribution of the 19-degree programmes across the four identified subpopulations relative to $k = 2, 3$

Subpopulation 1	Subpopulation 2	Subpopulation 3	Subpopulation 4
Early dropout (*k*=**2**)			
Aerospace Eng	Civil Eng	Chemical Eng	Biomedical Eng
Civil and Environmental Eng	Building Eng	Materials and Nanot. Eng	
Automation Eng	Telecom. Eng	Physics Eng	
Industrial Production Eng	Energy Eng	Mathematical Eng	
Electrical Eng	Management Eng		
Electronic Eng	Eng of Computing Systems		
Mechanical Eng			
Environ. and Land Planning Eng			
Late dropout (*k*=**3**)			
Biomedical Eng	Aerospace Eng	Civil Eng	Electronic Eng
Management Eng	Chemical Eng	Building Eng	Eng of Computing Systems
Mathematical Eng	Civil and Environmental Eng	Automation Eng	
Environ. and Land Planning Eng	Materials and Nanot. Eng	Telecom. Eng	
	Industrial Production Eng	Electrical Eng	
	Energy Eng		
	Physics Eng		
	Mechanical Eng		

For each k, the order of the four subpopulations is coherent to the one of the estimated random intercepts in Table 4

Acknowledgements The present research is part of the activities of "Dipartimento di Eccellenza 2023-2027".

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Data Availability Data available on request due to privacy restrictions.

Declarations

Ethics Approval This research does not contain any studies with humans or animals participation performed by any of the authors.

Conflict of Interest The authors declare no competing interests.

References

Agresti, A. (2018). *An introduction to categorical data analysis An introduction to categorical data analysis*. Wiley.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics, 55*(1), 117–128.

Azzimonti, L., Ieva, F., & Paganoni, A. M. (2013). Nonlinear nonparametric mixed-effects models for unsupervised classification Nonlinear nonparametric mixed-effects models for unsupervised classification. *Computational Statistics, 28*(4), 1549–1570.

Baum, C. F. (2016). Introduction to GSEM in Stata Introduction to gsem in stata. *ECON 8823: Applied Econometrics*.

Breslow, N. E., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika, 82*(1), 81–91.

Caliński, T. & Harabasz, J. (2013). *SAS/STAT® 13.1 User's Guide 13.1 user's guide*. SAS Institute Inc, Cary.

Cary, N. (2015). *SAS/STAT® 14.1 User's Guide*. Cary, NC: SAS Institute Inc.

Corp., I. (2021). IBM SPSS Statistics for Windows, Version 28.0 Ibm spss statistics for windows, version 28.0. Released 2021.

Daniels, M. J., & Gatsonis, C. (1997). Hierarchical polytomous regression models with applications to health services research Hierarchical polytomous regression models with applications to health services research. *Statistics in Medicine, 16*(20), 2311–2325.

De Leeuw, J., Meijer, E., & Goldstein, H. (2008). *Handbook of multilevel analysis*. Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Heagerty, P. J., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press.

Goldstein, H. (2011). *Multilevel statistical models* (vol. 922). John Wiley & Sons.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 159*(3), 505–513.

Hadfield, J. D., et al. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software, 33*(2), 1–22.

Hartzel, J., Agresti, A., & Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling, 1*(2), 81–102.

Hedeker, D., Gibbons, R., du Toit, M., & Cheng, Y. (2008). SuperMix: Mixed effects models. *Scientific Software International*.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine, 22*(9), 1433–1446.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Sage Publications, Inc.

King, G. (1989). *Unifying political methodology: The likelihood theory of statistical inference.* Cambridge University Press.

Kuss, O., & McLerran, D. (2007). A note on the estimation of the multinomial logistic model with correlated responses in SAS. *Computer Methods and Programs in Biomedicine, 87*(3), 262–269.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, 86–94.

Lindsay, B. G., et al. (1983). The geometry of mixture likelihoods, part II: The exponential family. *The Annals of Statistics, 11*(3), 783–792.

Long, J. S., & Long, J. S. (1997). *Regression models for categorical and limited dependent variables* (vol. 7). Sage.

Maggioni, A. (2020). *Semi-parametric generalized linear mixed effects model: An application to engineering BSc dropout analysis* (Unpublished doctoral dissertation).

Masci, C., Ieva, F., Agasisti, T., & Paganoni, A. M. (2021). Evaluating class and school effects on the joint student achievements in different subjects: A bivariate semiparametric model with random coefficients. *Computational Statistics*, 1–41.

Masci, C., Ieva, F., & Paganoni, A. M. (2022). Semiparametric multinomial mixed-effects models: A university students profiling tool. *The Annals of Applied Statistics, 16*(3), 1608–1632.

Masci, C., Paganoni, A. M., & Ieva, F. (2019). Semiparametric mixed effects models for unsupervised classification of Italian schools. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 182*(4), 1313–1342.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models* (wiley series in probability and statistics).

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86*(416), 899–909.

Pinheiro, J., & Bates, D. (2006). Mixed-effects models in S and S-PLUS. *Springer Science & Business Media*.

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. (https://www.R-project.org/)

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raudenbush, S. W. (2004). HLM 6: Hierarchical linear and nonlinear modeling. *Scientific Software International*.

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9*(1), 141–157.

Rights, J. D., & Sterba, S. K. (2016). The relationship between multilevel models and non-parametric multilevel mixture models: Discrete approximation of intraclass correlation, random coeffecient distributions, and residual heteroscedasticity. *British Journal of Mathematical and Statistical Psychology, 69*(3), 316–343.

Rodríguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 158*(1), 73–89.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *Winbugs user manual*. Citeseer.

Steele, F., Steele, F., Kallis, C., Goldstein, H., & Joshi, H. (2005). A multiprocess model for correlated event histories with multiple states, competing risks, and structural effects of one hazard on another. *Centre for Multilevel Modelling*: http://www.cmm.bristol.ac.uk/research/Multiprocess/mmcehmscrseoha.pdf.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas.*

Tutz, G., & Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics & Data Analysis, 22*(5), 537–557.

Wang, S., & Tsodikov, A. (2010). A self-consistency approach to multinomial logit model with random effects. *Journal of Statistical Planning and Inference, 140*(7), 1939–1947.

Zhao, Y., Staudenmayer, J., Coull, B. A., &Wand, M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science*, 35–51.