The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1608–1632
https://doi.org/10.1214/21-AOAS1559

SEMIPARAMETRIC MULTINOMIAL MIXED-EFFECTS MODELS: A UNIVERSITY STUDENTS PROFILING TOOL

BY CHIARA MASCI[a], FRANCESCA IEVA[b] AND ANNA MARIA PAGANONI[c]

MOX—Department of Mathematics, Politecnico di Milano, [a]*chiara.masci@polimi.it,* [b]*francesca.ieva@polimi.it,* [c]*anna.paganoni@polimi.it*

Many applicative studies deal with multinomial responses and hierarchical data. Performing clustering at the highest level of grouping, in multilevel multinomial regression, is also often of interest. In this study we analyse Politecnico di Milano data with the aim of profiling students, modelling their probabilities of belonging to different categories and considering their nested structure within engineering degree programmes. In particular, we are interested in clustering degree programmes standing on their effects on different types of student career. To this end, we propose an EM algorithm for implementing semiparametric mixed-effects models dealing with a multinomial response. The novel semiparametric approach assumes the random effects to follow a multivariate discrete distribution with an a priori unknown number of support points, that is, allowed to differ across response categories. The advantage of this modelling is twofold: the discrete distribution on random effects allows, first, to express the marginal density as a weighted sum, avoiding numerical problems in the integration step, typical of the parametric approach, and, second, to identify a latent structure at the highest level of the hierarchy where groups are clustered into subpopulations.

1. Introduction. The Italian higher education (HE) system measures a high level of dropout, with many students abandoning their studies during the Bachelor. According to the Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR), the dropout rate is around 28%, with almost two-thirds of students (20%) dropping out in the first two years (ANVUR (2018)). Understanding the dropout phenomenon is important and, in the last decades, is also receiving particular attention (Aina (2013), Aljohani (2016), Belloc, Maruotti and Petrella (2011)). Many studies aim at individuating personal features of students who are more likely to drop out in order to partially prevent the phenomenon (Aljohani (2016)).

The *student profile for enhancing tutoring engineering* (SPEET) project is an ERASMUS$^+$ project aimed at extracting useful information from academic data, provided by its partners,[1] and determining and categorizing different profiles of engineering students across Europe (Barbu et al. (2019)). The essence of the SPEET project is to apply data mining algorithms in order to extract information about students, to profile them identifying students at-risk of dropout and to define tutoring activities to support them (De Freitas et al. (2015)).

This work is within the SPEET project and focuses on the analysis of data about Politecnico di Milano (PoliMI) students attending Bachelor courses in engineering. The PoliMI dropout rate in engineering is around 30%, with the majority of students dropping out during the first year (Cannistrà et al. (2021)). This high percentage represents a net waste of

Received September 2020; revised June 2021.

Key words and phrases. Semiparametric statistics, multinomial mixed-effects regression, unsupervised clustering, higher education.

[1]SPEET consortium is composed of six European universities: Universitat Autònoma de Barcelona (UAB)—Barcelona, Spain; Instituto Politecnico de Braganca (IPB)—Braganca, Portugal; Opole University of Technology—Opole, Poland; Politecnico di Milano (PoliMI)—Milano, Italy; Universidad de Leòn—Leòn, Spain; University of Galati *Dunarea de Jos*—Galati, Romania.

resources, since educating students is a costly activity that generates returns in the long term. Among students who abandon their studies, we observe students who abandon their university career at its very beginning (due, e.g., to a wrong choice of the faculty or to an initial discouragement) and students who abandon after a while. In this perspective, concluded careers of students can be classified as *graduate*, *early dropout* (i.e., careers concluded with a dropout within the first three semesters since the enrolment) and *late dropout* (i.e., careers concluded with a dropout after more than three semesters since the enrolment). Studies confirm that the causes and dynamics of these two types of dropout might be different, and their distinction and identification is essential in the perspective of supporting students at-risk (Cannistrà et al. (2021)).

PoliMI offers about 20 different engineering degree programmes and students are nested within them. Degree programmes have heterogeneous internal dynamics, students characteristics and study plans; these aspects might lead to different dropout rates and motivations. In this study we aim at profiling engineering students, modelling their probabilities of belonging to different categories, standing on their personal and early career information and considering the degree programme the student is attending. In particular, we are also interested in identifying latent subpopulations of degree programmes standing on their effects on different types of student career. To this end, we develop a semiparametric multinomial mixed-effects model, whose random effects follow a discrete distribution with an unknown number of mass points. This modelling allows to identify a latent structure at the highest level of the hierarchy, where groups (i.e., degree programmes) are clustered into subpopulations, standing on their effect on the outcome, that is, the probability of students to belong to different categories of the multinomial response variable.

Many studies deal with hierarchical data, that is, data containing observations naturally nested within groups. Examples of such data are longitudinal data, repeated measurements for each subject in a study, or multilevel data (e.g., patients nested within hospitals or students nested within schools). One of the most common approaches for modelling them are mixed-effects models, that are regression or classification models that include in the linear predictor both *fixed effects* associated to the entire population and *random effects* associated to the groups, drawn at random from the population, in which observations are nested (Goldstein (2011)). This mechanism allows to account for correlation structures among the nested observations, which are not independent, modelling the within-group correlation.

Typically, mixed-effects linear models assume both the random effects and the errors to follow a Gaussian distribution, and these models are intended for grouped data in which the response variable is continuous (Pinheiro and Bates (2006)). When the response has a different distribution in the exponential family, generalized linear mixed-effects models (GLMMs) extend generalized linear models to include random effects (Diggle et al. (2002), Agresti (2018)). In GLMMs the response distribution is defined, conditionally, on the random effects that are usually assumed to be multivariate normal. Under this assumption the marginal distribution of the response can be obtained by integrating out the random effects, but it does not have closed form. In order to approximate the marginal density, various numerical methods have been proposed in the literature: numerical integration using Gauss–Hermite quadrature (e.g., Anderson and Aitkin (1985)), Monte Carlo techniques (e.g., Booth and Hobert (1999), McCulloch (1994, 1997)) or approximation methods, such as Laplace approximation and Taylor series expansions (e.g., Breslow and Clayton (1993), Wolfinger and O'connell (1993)).

Although GLMMs have been developed for a consistent set of response distributions in the exponential family (among the others, binomial, Poisson, Gamma, Inverse Gaussian), there has been less development for a multinomial response. In particular, the majority of the research in this area focuses on ordinal models with logit and probit link functions for

cumulative probabilities (Anderson, Kim and Keller (2013), Coull and Agresti (2000), Dos Santos and Berridge (2000)), while nominal responses have received less attention, probably due to the higher level of complexity they require. Indeed, an appropriate link function for nominal responses is the baseline-category logit, where fixed and random coefficients vary according to the response category. Considering a multinomial response assuming K different categories, the baseline-category logit approach implies the presence of $K-1$ vectors of fixed effects coefficients and $K-1$ random effects coefficients distributions. Mixed-effects linear models for a multinomial response are, therefore, often treated as multivariate models, where the integration issues typical of GLMMs grow in complexity (De Leeuw, Meijer and Goldstein (2008)). Various approximations for evaluating the integral over the random effects distribution have been proposed in the literature: the most frequently used methods are based on first- or second-order Taylor expansions (Goldstein and Rasbash (1996)), on a combination of a fully multivariate Taylor expansion and a Laplace approximation (Raudenbush, Yang and Yosef (2000)) or using Gauss–Hermite quadrature (Stroud and Secrest (1966)). Regarding the random effects, they can be estimated using empirical Bayes methods (Bock and Aitkin (1981)). Nonetheless, these cited procedures are computationally very complex (McCulloch and Searle (2001)) and many authors have reported biased estimates using them (Breslow and Lin (1995), Raudenbush, Yang and Yosef (2000), Rodríguez and Goldman (1995)). Moreover, specific softwares have been developed to perform these kind of estimates (among the others, HLM (Raudenbush (2004)), MLwiN (Steele et al. (2005)) and WinBugs (Spiegelhalter et al. (2003))), but they resulted to be not very flexible, and they often require a high level of expertise on behalf of the user. In one of the most recent works (Hadfield et al. (2010)), the authors develop a Markov chain Monte Carlo (MCMC) method for multiresponse 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 context, where the distinction between fixed and random effects does not technically exist, and the user should define the priors on the parameters. If the priors are not defined (and, therefore, default priors are used) or are improperly defined, this can lead to both inferential and numerical problems. The relative *MCMCglmm* R package (Hadfield et al. (2010)) is, to the best of our knowledge, the only R package (R Core Team (2019)) that performs parametric mixed-effects multinomial regression.

In this paper, guided by the aim of modelling a categorical response variable, with $K > 2$ categories and, taking into account the nested structure of data, of performing clustering at the highest level of grouping, we propose a semiparametric mixed-effects linear model for a multinomial response that consists in a novel approach in which random effects coefficients, instead of being multivariate normal, have a multivariate discrete distribution with an a priori unknown number of support points. In particular, considering a multinomial response assuming K different categories and the baseline-category logit approach, each one of the $K-1$ logits is identified by a specific random effects coefficients distribution with an unknown finite number of support points, that is, allowed to differ across logits. This approach is inspired by the work proposed in Masci, Paganoni and Ieva (2019), where the authors propose a semiparametric mixed-effects model where random coefficients follow a discrete distribution but for a continuous response. This work has connections with the literature regarding growth mixture models (GMMs) and latent class growth models (LCGMs) (see Heinen (1996), McCulloch et al. (2002), Muthén (2004), Nagin (1999) for discussion). GMMs and LCGMs have been broadly applied in the context of social sciences (Shaw, Lacourse and Nagin (2005), Shaw et al. (2003), Nagin et al. (2018)). They are used to model longitudinal data to estimate the average growth, the amount of variation across individuals in growth intercept and slopes and the influence of covariates on this variation. They allow for the existence of latent trajectory classes where groups of individual growth trajectories

vary around different behaviours. The main difference between these methods and our proposed one is that GMMs and LCGMs need to fix a priori the number of latent classes to be identified, while, in our approach, we estimate it together with the other model parameters. In the educational field as well as in other social sciences, defining a priori the number of latent behaviours at the highest level of grouping is not a trivial task and having at disposal a way to estimate it standing on different criteria might consist in a great improvement. Moreover, being GMMs and LCGMs intended for longitudinal data, the set of time instants in which dependent and independent variables are evaluated is the same across groups or individuals, meaning that covariates are fixed across groups or individuals.

The advantage introduced by the proposed modelling is twofold: (i) the former is that, by assuming a discrete distribution at the highest level of the hierarchy, we avoid the integration issues relative to the continuous distribution; (ii) the latter is that this assumption allows to identify a latent structure within the highest level of the hierarchy, that is, the presence of subpopulations among groups. Moreover, this modelling allows to investigate how the latent structure at the highest level of the hierarchy does change across categories with respect to the baseline. To estimate the semiparametric model parameters, we propose an expectation-maximization (EM) algorithm (Dempster, Laird and Rubin (1977)) that alternates the estimates of fixed effects and random effects, until the convergence is reached. A similar approach for a multinomial response has been proposed by Hartzel, Agresti and Caffo (2001), where the authors generalize Aitkin (1999) for an ordinal random effects model, treating the random effects in a nonparametric manner and assuming them to follow a discrete distribution. Although there are many parallels with this work, the authors in Hartzel, Agresti and Caffo (2001) consider the only case of a random intercept (i.e., not considering random effects covariates), and they need to specify a priori the number of support points of the random effects distribution.

When modelling student different types of dropout, this method allows, first, to refine the classification of student careers, with respect to the simple binary classificator *dropout* vs. *graduate*, considering multiple categories at once, and second, to identify a latent structure among degree programmes that is response-specific, revealing how the effects of the degree programmes do change with respect to different student profiles. In addition, the a posteriori eventual profiling of degree programmes subpopulations represents an easy tool to identify potential drivers of different students dropout trends across courses.

After presenting the semiparametric mixed-effects model for a multinomial response and the EM algorithm—called *MSPEM algorithm*—to estimate its parameters, we show a simulation study and, lastly, our case study in which we apply the algorithm to Politecnico di Milano data for modelling university student dropout, comparing its results with the ones obtained by applying the MCMCglmm algorithm. The paper is organised as follows: in Section 2 we present the semiparametric mixed-effects model for a multinomial response and the MSPEM algorithm; in Section 3 we present a simulation study testing the algorithm within different settings; in Section 4 we apply the MSPEM algorithm for modelling student dropout, and we compare the results with the ones of the parametric MCMCglmm algorithm; in Section 5 we draw our conclusions and discuss some future perspectives.

2. Methodology: Semiparametric mixed-effects model for a multinomial response. Let's consider a multinomial logistic regression model for nested data with a two-level hierarchy (Agresti (2018), De Leeuw, Meijer and Goldstein (2008)), where each observation j, for $j = 1, \dots, n_i$, is nested within a group i, for $i = 1, \dots, I$. Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})$ be the n_i-dimensional response vector for observations within the ith group. The multinomial

distribution with K categories relative to Y_{ij} is the following:

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

where $k = 1, \dots, K$ are the 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 $Y_{ij} = k$, that is, π_{ijk}, is given by

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

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) of the fixed effects, $\boldsymbol{\alpha}_k$ is the $p \times 1$ vector of regression parameters of the fixed effects, $\mathbf{z}_{ij}$ is the $q \times 1$ covariates vector of the random effects (includes a 1 for the intercept) and $\boldsymbol{\delta}_{ik}$ is the $q \times 1$ vector of regression parameters of the random effects. In this formulation we model $K - 1$ contrasts, between each category k, for $k = 2, \dots, K$, and the reference category,[2] 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, and we estimate the parameters (for both fixed and random effects) relative to the $(K - 1)$ contrasts. Let us observe that, starting from equation (2), the log-odds of each response with respect to the reference category are

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

Logit models for nominal response basically pair each category with a baseline category. We, therefore, observe $K - 1$ contrasts, where each contrast k', $k' = 1, \dots, K - 1$, is characterized by the set of *contrast-specific* parameters ($\boldsymbol{\alpha}_{k'}$; $\boldsymbol{\delta}_{ik'}$, for $i = 1, \dots, I$) 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).[3] For each contrast the *contrast-specific* parameters describe the latent structure at the higher level of the hierarchy.

In order to set the parameters estimation procedure, we need to model the probability of Y_{ij} conditional on the random effects distribution. In particular, considering $\mathbf{A} = (\boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_K)$ and $\boldsymbol{\Delta}_i = (\boldsymbol{\delta}_{i2}, \dots, \boldsymbol{\delta}_{iK})$, the conditional distribution of Y_{ij} takes the following form:

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

[2] We consider "1" as the reference category, but this choice is arbitrary, and it does not affect the model formulation.

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

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

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

In a parametric framework, $\boldsymbol{\delta}_{ik}$ are usually assumed to follow a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Omega_k)$ (De Leeuw, Meijer and Goldstein (2008)). To standardize the multiple random effects, we can decompose $\boldsymbol{\delta}_{ik}$ as $\boldsymbol{\delta}_{ik} = T_k\boldsymbol{\theta}_i$, where $T_k T_k'$ is the Cholesky decomposition of Ω_k and $\boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{0}, I)$ (De Leeuw, Meijer and Goldstein (2008)). T_k is the random effects variance term. Given this formulation, the marginal density of $\mathbf{Y}_i$, $h(\mathbf{Y}_i)$, is expressed as the integral of the conditional likelihood, $p(\mathbf{Y}_i|\boldsymbol{\theta})$, weighted by the prior density $g(\boldsymbol{\theta})$,

$$
h(\mathbf{Y}_i) = \int_{\boldsymbol{\theta}} p(\mathbf{Y}_i|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{6}
$$

where $g(\boldsymbol{\theta})$ is the multivariate standard normal density. To obtain the maximum likelihood estimates of the parameters, the marginal log-likelihood from the I level-2 units, $\log L = \sum_{i=1}^{I} \log h(\mathbf{Y}_i)$, can be maximized, but it implies many computational issues (Skrondal and Rabe-Hesketh (2004)). Indeed, integration over the random effects distribution must be performed and this is often intractable.

Following the approach presented in Masci, Paganoni and Ieva (2019), we move to a semi-parametric framework, assuming the coefficients of the random effects to follow a discrete distribution with an a priori unknown number of support points. Under this assumption the multinomial logit takes the form

$$
\eta_{ijk} = \mathbf{x}_{ij}'\boldsymbol{\alpha}_k + \mathbf{z}_{ij}'\mathbf{b}_{m_k k} \quad m_k = 1, \ldots, M_k, k = 2, \ldots, K, \tag{7}
$$

where M_k is the total number of support points of the discrete distribution of $\mathbf{b}$ relative to the kth category, for $k = 2, \ldots, K$. The random effects distribution relative to each category k, for $k = 2, \ldots, K$, can be expressed as a set of points $(\mathbf{b}_{1k}, \ldots, \mathbf{b}_{M_k k})$, where $M_k \leq I$ and $\mathbf{b}_{m_k k} \in \mathcal{R}^q$ for $m_k = 1, \ldots, M_k$, and a set of weights $(w_{1k}, \ldots, w_{M_k k})$, where $\sum_{m_k=1}^{M_k} w_{m_k k} = 1$ and $w_{m_k} \geq 0$,

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

We call P_k^* the discrete distribution relative to each $k = 2, \ldots, K$ that belongs to the class of all probability measures on $\mathcal{R}^q$. 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 (7). In particular, $w_{m_k k} = P(\boldsymbol{\delta}_{ik} = \mathbf{b}_{m_k k})$, for $i = 1, \ldots, I$. The maximum likelihood estimator $\hat{P}_k^*$ of P_k^* can be obtained following the theory of mixture likelihoods in Lindsay (1983a), Lindsay (1983b), 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. In particular, Lindsay (1983a), Lindsay (1983b) faced statistical problems (existence, discreteness, support size characterization and uniqueness), transforming them in geometrical problems, concerning support hyperplanes of the convex hull of the likelihood curve.

2.1. *EM algorithm for semiparametric mixed-effects model for a multinomial response*. Given the proposed formulation, we implement an EM algorithm for the joint estimations of $\boldsymbol{\alpha}_k$, $(\mathbf{b}_{1k}, \ldots, \mathbf{b}_{M_k k})$ and $(w_{1k}, \ldots, w_{M_k k})$, for $k = 2, \ldots, K$, which is performed through the maximization of the likelihood, mixture by the discrete distribution of the random effects. Under these assumptions the marginal likelihood can be obtained as a weighted sum of all the conditional probabilities. In the extreme case of $K = 2$, that is, a classical logistic regression, we would have a unique distribution of $\mathbf{b}$, that is $(\mathbf{b}_1, \ldots, \mathbf{b}_M)$ with weights $(w_1, \ldots, w_M)$, and the marginal likelihood of $\mathbf{Y}_i$ would take the form

$$h(\mathbf{Y}_i|\boldsymbol{\alpha}) = \sum_{m=1}^{M} w_m p(\mathbf{Y}_i|\boldsymbol{\alpha}, \mathbf{b}_m). \tag{8}$$

In the case of a generic $K > 2$, the likelihood in equation (8) generalizes to the weighted sum of the likelihood of $\mathbf{Y}$, conditioned to all the possible combinations, that are $M_{tot} = \prod_{k=2}^{K} M_k$ of the values of the $(K-1)$ discrete distributions of random effects. We can write this likelihood as

$$h(\mathbf{Y}_i|\boldsymbol{A}) = \sum_{m=1}^{M_{\text{tot}}} w_m p(\mathbf{Y}_i|\boldsymbol{A}, \mathbf{B}_m), \tag{9}$$

where w_m is the weight relative to the mth combination of the $(K-1)$ weights relative to the $(K-1)$ contrasts and, analogously, $\mathbf{B}_m$ is the mth combination of the random effects coefficients relative to the $(K-1)$ contrasts. Assuming the independence of the random effects distributions across the $(K-1)$ contrasts, we can marginalize the weights and write the likelihood as follows:

$$\begin{aligned} &h(\mathbf{Y}|\boldsymbol{A}) \\ &\quad = w_{12} \times w_{13} \times \cdots \times w_{1K} \times p(\mathbf{Y}|\boldsymbol{A}, \mathbf{b}_{12}, \mathbf{b}_{13}, \ldots, \mathbf{b}_{1K}) \\ &\qquad + w_{22} \times w_{13} \times \cdots \times w_{1K} \times p(\mathbf{Y}|\boldsymbol{A}, \mathbf{b}_{22}, \mathbf{b}_{13}, \ldots, \mathbf{b}_{1K}) \\ &\qquad + \cdots \\ &\qquad + \cdots \\ &\qquad + w_{M_2 2} \times w_{M_3 3} \times \cdots \times w_{M_K K} \times p(\mathbf{Y}|\boldsymbol{A}, \mathbf{b}_{M_2 2}, \mathbf{b}_{M_3 3}, \ldots, \mathbf{b}_{M_K K}). \end{aligned} \tag{10}$$

The EM algorithm proposed is an iterative algorithm that alternates two steps: 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 $j = 1, \ldots, n_i$ and $i = 1, \ldots, I$. The algorithm allows the number n_i, for $i = 1, \ldots, I$, of observations to be different across groups, but within each group missing data are not handled. At each iteration the EM algorithm updates the parameters to increase the likelihood in equation (10), and it continues until convergence or until a fixed number of iterations is reached. In particular, the update for the parameters relative to each response category k, for $k = 2, \ldots, K$, is given by

$$w_{m_k k}^{(\text{up})} = \frac{1}{I} \sum_{i=1}^{I} W_{i m_k k} \quad m_k = 1, \ldots, M_k, \tag{11}$$

where

$$W_{i m_k k} = \frac{w_{m_k k} p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{b}_{m_k k}, \bar{\mathbf{b}}_{l \neq k})}{\sum_{\gamma=1}^{M_k} w_{\gamma k} p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{b}_{\gamma k}, \bar{\mathbf{b}}_{l \neq k})} \quad m_k = 1, \ldots, M_k, \tag{12}$$

and

$$(\boldsymbol{\alpha}_k^{(\mathrm{up})}, \mathbf{b}_{1k}^{(\mathrm{up})}, \ldots, \mathbf{b}_{M_k k}^{(\mathrm{up})}) = \underset{\boldsymbol{\alpha}_k, \mathbf{b}_{m_k k}}{\arg\max} \sum_{m_k=1}^{M_k} \sum_{i=1}^{I} W_{i m_k k} \times \ln p\big(\mathbf{y}_i | \boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{l\neq k}^{(\mathrm{old})}, \mathbf{b}_{m_k k}, \bar{\mathbf{b}}_{l\neq k}^{(\mathrm{old})}\big). \tag{13}$$

In equation (13) the random effects coefficients $\bar{\mathbf{b}}_l^{(\mathrm{old})}$, for $l \neq k$, are the mean of the discrete distribution relative to the lth category, $\bar{\mathbf{b}}_l^{(\mathrm{old})} = \sum_{m_l=1}^{M_l} w_{m_l l}^{(\mathrm{old})} \mathbf{b}_{m_l l}^{(\mathrm{old})}$, computed in the previous iteration. In particular,

$$p\big(\mathbf{y}_i | \boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{l\neq k}^{(\mathrm{old})}, \mathbf{b}_{m_k k}, \bar{\mathbf{b}}_{l\neq k}^{(\mathrm{old})}\big) = \prod_{j=1}^{n_i} \prod_{\gamma=1}^{K} \left(\frac{\exp(\eta_{ij\gamma})}{1 + \sum_{\nu=2}^{K} \exp(\eta_{ij\nu})} \right)^{\{\mathbf{1}_{y_{ij}=\gamma}\}}, \tag{14}$$

where

$$\eta_{ij\gamma} = \begin{cases} \mathbf{x}_{ij}' \boldsymbol{\alpha}_k + \mathbf{z}_{ij}' \mathbf{b}_{m_k} & \text{if } \gamma = k, \\ \mathbf{x}_{ij}' \boldsymbol{\alpha}_\gamma^{(\mathrm{old})} + \mathbf{z}_{ij}' \sum_{m_\gamma=1}^{M_\gamma} w_{m_\gamma \gamma}^{(\mathrm{old})} \mathbf{b}_{m_\gamma \gamma}^{(\mathrm{old})} & \text{if } \gamma \neq k. \end{cases} \tag{15}$$

The weight $w_{m_k k}^{(\mathrm{up})}$ is the mean over the I groups of their weights relative to the m_kth subpopulation, relative to category k. Coefficient $W_{i m_k k}$ represents the probability that group i belongs to subpopulation m_k conditionally on observations $\mathbf{y}_i$ and, given the fixed coefficients $\boldsymbol{A}$, with respect to category k. The maximization step in equation (13) involves two steps, and it is done iteratively. In the first step, for each category k, for $k = 2, \ldots, K$, we compute arg max with respect to the support points $\mathbf{b}_{m_k k}$, for $m_k = 1, \ldots, M_k$, keeping $\boldsymbol{A}$ and $\bar{\mathbf{b}}_l$, for $l \neq k$, fixed to the values computed in 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, that is,

$$\mathbf{b}_{m_k k}^{(\mathrm{up})} = \underset{\mathbf{b}_k}{\arg\max} \sum_{i=1}^{I} W_{i m_k k} \ln p(\mathbf{y}_i | \boldsymbol{A}, \mathbf{b}_k, \bar{\mathbf{b}}_{l\neq k}) \quad m_k = 1, \ldots, M_k, k = 2, \ldots, K. \tag{16}$$

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 equation (13) with respect to $\boldsymbol{\alpha}_k$.

Since $w_{m_k k} = P(\boldsymbol{\delta}_{ik} = \mathbf{b}_{m_k k})$, then

$$\begin{aligned} W_{i m_k k} &= \frac{w_{m_k k} p(\mathbf{y}_i | \boldsymbol{A}, \mathbf{b}_{m_k k}, \bar{\mathbf{b}}_{l\neq k})}{\sum_{\gamma=1}^{M_k} w_{\gamma k} p(\mathbf{y}_i | \boldsymbol{A}, \mathbf{b}_{\gamma k}, \bar{\mathbf{b}}_{l\neq k})} \\ &= \frac{p(\boldsymbol{\delta}_{ik} = \mathbf{b}_{m_k k}) p(\mathbf{y}_i | \boldsymbol{A}, \mathbf{b}_{m_k k}, \bar{\mathbf{b}}_{l\neq k})}{p(\mathbf{y}_i | \boldsymbol{A})} \\ &= \frac{p(\mathbf{y}_i, \boldsymbol{\delta}_{ik} = \mathbf{b}_{m_k k} | \boldsymbol{A}, \bar{\mathbf{b}}_{l\neq k})}{p(\mathbf{y}_i | \boldsymbol{A}, \bar{\mathbf{b}}_{l\neq k})} \\ &= p(\boldsymbol{\delta}_{ik} = \mathbf{b}_{m_k k} | \mathbf{y}_i, \boldsymbol{A}, \bar{\mathbf{b}}_{l\neq k}). \end{aligned} \tag{17}$$

Therefore, to compute the point $\mathbf{b}_{m_k k}$ for each group i, for $i = 1, \ldots, I$, we maximize the conditional probability of $\boldsymbol{\delta}_{ik}$, given the observations $\mathbf{y}_i$, the coefficient $\boldsymbol{A}$ and the random effects relative to the other categories l, $l \neq k$. The estimates of the coefficients $\boldsymbol{\delta}_{ik}$ of the

random effects for each group and each category is obtained by maximizing W_{im_kk} over m_k, that is,

$$\hat{\hat{\boldsymbol{\delta}}}_{ik} = \mathbf{b}_{\tilde{m}_k k} \quad \text{where } \tilde{m}_k = \arg\max_{m_k} W_{im_kk},$$

$$i = 1, \ldots, N, \quad k = 2, \ldots, K. \tag{18}$$

Supplementary Material A in Masci, Ieva and Paganoni (2022) reports the increasing likelihood property proof of this parameters update procedure.

During the iterations of the EM algorithm, we perform the reduction of the support points of the random effects discrete distributions in order to identify, for $k = 2, \ldots, K$, $M_k* < I$ subpopulations that describe the latent structure relative to each contrast $k' = k - 1$. To this end, we fix a threshold distance D_k, for $k = 2, \ldots, K$, and when two mass points, relative to category k, $\mathbf{b}_{m_kk}$ and $\mathbf{b}_{m_lk}$ are closer than D_k, they collapse to a unique point $\mathbf{b}_{m_{k,l}k} = (\mathbf{b}_{m_kk} + \mathbf{b}_{m_lk})/2$ with associated weight $w_{m_{l,k}k} = w_{m_lk} + w_{m_kk}$. The threshold D_k is allowed to differ across the categories, that is, we may choose $(K - 1)$ different values, one for each of the $(K - 1)$ random effects distributions, depending on the problem. For each category k, $k = 2, \ldots, K$, the first two masses that collapse to a unique point are the two masses with the minimum Euclidean distance, among the couples of masses with Euclidean distance less than D_k, and so on.

An other criterion for the support reduction regards the minimum number of groups within each subpopulation. Starting from a given iteration up to the end, we fix a threshold $\tilde{w}_k$, for $k = 2, \ldots, K$, and we remove mass points with weight $w_{m_kk} < \tilde{w}_k$. This criterion goes in the direction of the outlier detection, since the groups that will not be associated to any subpopulation with a minimum weight $\tilde{w}_k$, for $k = 2, \ldots, K$, will result as anomalous groups. By tuning the two parameters D_k and $\tilde{w}_k$, we can govern these two collapsing criteria. Supplementary Material B in Masci, Ieva and Paganoni (2022) reports further insights about the discrete distribution support points initialization, the support points collapse criteria and the convergence criteria.

Some final issues that deserve attention regard inference and stability of the parameter estimates and model identifiability. Theoretically, the asymptotic inferential theory for the fixed effects estimation would parallel the standard maximum likelihood theory, but this theory is partly lacking because of the unknown mixture support size. Despite this, Hartzel (2000) examined the Wald and likelihood-ratio tests for mixed-effects models for a multinomial response, concluding that they provide appropriate inference for the semiparametric maximum likelihood approach. Moreover, regarding the stability of fixed effects parameters, studies on the comparison between parametric and nonparametric approach confirm that, for a semiparametric approach, the estimated bias for $\boldsymbol{A}$ is similar to the parametric approach one. In particular, they suggest that parametric and semiparaemtric approach produce essentially unbiased estimates of $\boldsymbol{A}$, with similar behavior under various random effects distributions and subpopulations sizes (Hartzel, Agresti and Caffo (2001)). Regarding identifiability issues, a mixture is identifiable if it is uniquely characterized, that is, if two distinct sets of parameters defining the mixture can not yield to the same distribution. Again, Hartzel (2000) provided sufficient conditions for the identifiability of overdispersed multinomial regression models, but we are aware that further studies are needed for the more general case considered here.

3. Simulation study. In this section we propose a simulation study to test the MSPEM algorithm performance under different settings. Let consider a categorical response variable that assumes three possible values in $\mathcal{K} = \{1, 2, 3\}$, where $k = 1$ is the reference category. We simulate three different settings: (i) one considering only a random intercept, (ii) one

considering only a random slope and (iii) one considering both random intercept and random slope. In the application for modelling student dropout, shown in Section 4, we have a three-categories response, and we consider the case of only a random intercept. Here in the simulation study, we maintain the three-categories response to ease the reader, and, besides the case (i) of a random intercept, we add the other two random effects cases in order to show the method results in more complex settings. We include two covariates in the model (considered as both fixed or one random and one fixed) to be in line with the case study.

We consider $I = 100$ groups of data, where each group contains 200 observations,[4] and we induce the presence of three subpopulations regarding category $k = 2$, that is, $M_2 = 3$, and two subpopulations regarding category $k = 3$, that is, $M_3 = 2$. In particular, for $j = 1, \dots, 200$ and $i = 1, \dots, 100$, we consider the model

$$\begin{aligned} \pi_{ijk} &= P(Y_{ij} = k) = \frac{\exp(\eta_{ijk})}{1 + \sum_{l=2}^{3} \exp(\eta_{ijl})} \quad \text{for } k = 2, 3; \\ \pi_{ij1} &= P(Y_{ij} = 1) = \frac{1}{1 + \sum_{l=2}^{3} \exp(\eta_{ijl})}, \end{aligned} \tag{19}$$

where the linear predictor $\boldsymbol{\eta}_{ik} = (\eta_{i1k}, \dots, \eta_{i200k})$ is defined by the following data generating process[5] (DGP):

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

$$\begin{aligned} \boldsymbol{\eta}_{i2} &= \begin{cases} +4\mathbf{x}_{1i} - 3\mathbf{x}_{2i} - 7 & i = 1, \dots, 30, \\ +4\mathbf{x}_{1i} - 3\mathbf{x}_{2i} - 4 & i = 31, \dots, 60, \\ +4\mathbf{x}_{1i} - 3\mathbf{x}_{2i} - 2 & i = 61, \dots, 100, \end{cases} \\ \boldsymbol{\eta}_{i3} &= \begin{cases} -2\mathbf{x}_{1i} + 2\mathbf{x}_{2i} - 5 & i = 1, \dots, 60, \\ -2\mathbf{x}_{1i} + 2\mathbf{x}_{2i} - 2 & i = 61, \dots, 100, \end{cases} \end{aligned} \tag{20}$$

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

$$\begin{aligned} \boldsymbol{\eta}_{i2} &= \begin{cases} -1 - 3\mathbf{x}_{1i} + 5\mathbf{z}_{1i} & i = 1, \dots, 30, \\ -1 - 3\mathbf{x}_{1i} + 2\mathbf{z}_{1i} & i = 31, \dots, 60, \\ -1 - 3\mathbf{x}_{1i} - 1\mathbf{z}_{1i} & i = 61, \dots, 100, \end{cases} \\ \boldsymbol{\eta}_{i3} &= \begin{cases} -2 + 2\mathbf{x}_{1i} - 2\mathbf{z}_{1i} & i = 1, \dots, 60, \\ -2 + 2\mathbf{x}_{1i} - 6\mathbf{z}_{1i} & i = 61, \dots, 100, \end{cases} \end{aligned} \tag{21}$$

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

$$\begin{aligned} \boldsymbol{\eta}_{i2} &= \begin{cases} -5\mathbf{x}_{1i} - 6 + 5\mathbf{z}_{1i} & i = 1, \dots, 30, \\ -5\mathbf{x}_{1i} - 4 + 2\mathbf{z}_{1i} & i = 31, \dots, 60, \\ -5\mathbf{x}_{1i} - 8 - 1\mathbf{z}_{1i} & i = 61, \dots, 100, \end{cases} \\ \boldsymbol{\eta}_{i3} &= \begin{cases} +2\mathbf{x}_{1i} + 1 - 4\mathbf{z}_{1i} & i = 1, \dots, 60, \\ +2\mathbf{x}_{1i} - 1 + 2\mathbf{z}_{1i} & i = 61, \dots, 100, \end{cases} \end{aligned} \tag{22}$$

[4]The number of observations is allowed to be different across groups. Here, to facilitate the reader and without loss of generality, we consider it unvaried across groups.

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

TABLE 1

MSPEM algorithm performance across the 500 *runs for each of the three cases. First column reports the number of runs out of* 500 *in which the algorithm identifies the correct number of subpopulations that are simulated in the DGP in equations* (20), (21) *and* (22); *second column reports the number of runs out of the number of runs in which the algorithm identifies* $M_2 = 3$ *and* $M_3 = 2$ (*reported in the first column*) *in which the algorithm correctly assigns each group to the correspondent subopulation*

	# runs in which MSPEM identifies $M_2 = 3$ and $M_3 = 2$	# runs in which MSPEM correctly classifies all groups into subpopulations
(i) Random intercept case	473 out of 500	470 out of 473
(ii) Random slope case	453 out of 500	427 out of 453
(iii) Random intercept and slope case	422 out of 500	315 out of 422

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.

We perform 500 runs of the MSPEM algorithm for each of the three settings shown in (20), (21) and (22). We fix $D_k = 1$ for $k = \{2, 3\}$ as threshold value for the support points collapse criterium, $\tilde{w}_k = 0$ for $k = \{2, 3\}$ to allow the presence of subpopulations containing a single group, `tollR` $=$ `tollF` $= 0.01$, `itmax` $= 50$ and `it1` $= 30$ (see Supplementary Material B in Masci, Ieva and Paganoni (2022)). The algorithm starts by considering $M_k^* = 20$, for $k = \{2, 3\}$ support points, extracted from a uniform distribution with support on the entire range of possible values, that is, the range of the distribution of coefficients obtained by fitting I distinct multinomial logistic regressions. Starting weights are uniformly distributed on these 20 support points (see Supplementary Material B in Masci, Ieva and Paganoni (2022) for more details). In all the runs, the algorithm converges in a number of iterations that range between *five* and 10. Table 1 reports the number of runs out of 500 in which the algorithm identifies the simulated number of subpopulations (i.e., $M_2 = 3$ and $M_3 = 2$) and correctly assigns groups to the identified subpopulations for all the three settings.

The algorithm correctly identifies the simulated number of subpopulations in more than 84% of the runs and classifies all groups into the correspondent subpopulations in more than 63% of the runs for all of the three cases. In the remaining runs the algorithm usually identifies a higher number of subpopulations (i.e., M_2 equal to 4,, instead of 3, and M_3 equal to 3 instead of 2), being it more sensitive to the variability among the data or misclassifies a very small percentage of groups into the identified subpopulations (usually no more than *three* groups out of 100).

Table 2 reports the results of the estimated coefficients in the three different settings. Descriptive statistics about estimated fixed effects coefficients are computed on the total number of runs, while random effects ones are computed only on the runs where the estimated number of subpopulations corresponds to the simulated one (i.e., the majority of the cases). Indeed, 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 coefficients estimates do not result to be affected by the number of subpopulations identified in the data. The estimated coefficients are very close to the original ones, and their variability is low. The identification of supopulations and their relative numerosity depends on the tuning parameter D_k, that, given the order of magnitude of the simulated coefficients, we fix equal to the unit ($D_k = 1$, for $k = \{2, 3\}$). Increasing the value of D_k, the mass points of the random effects coefficients distribution that have higher distances will collapse to a unique point, and the MSPEM algorithm will be less sensitive to the variability among the I groups, identifying a smaller number of subpopulations. On the opposite, decreasing the value of D_k, mass points that have smaller distances (not smaller than D_k), will not collaspe

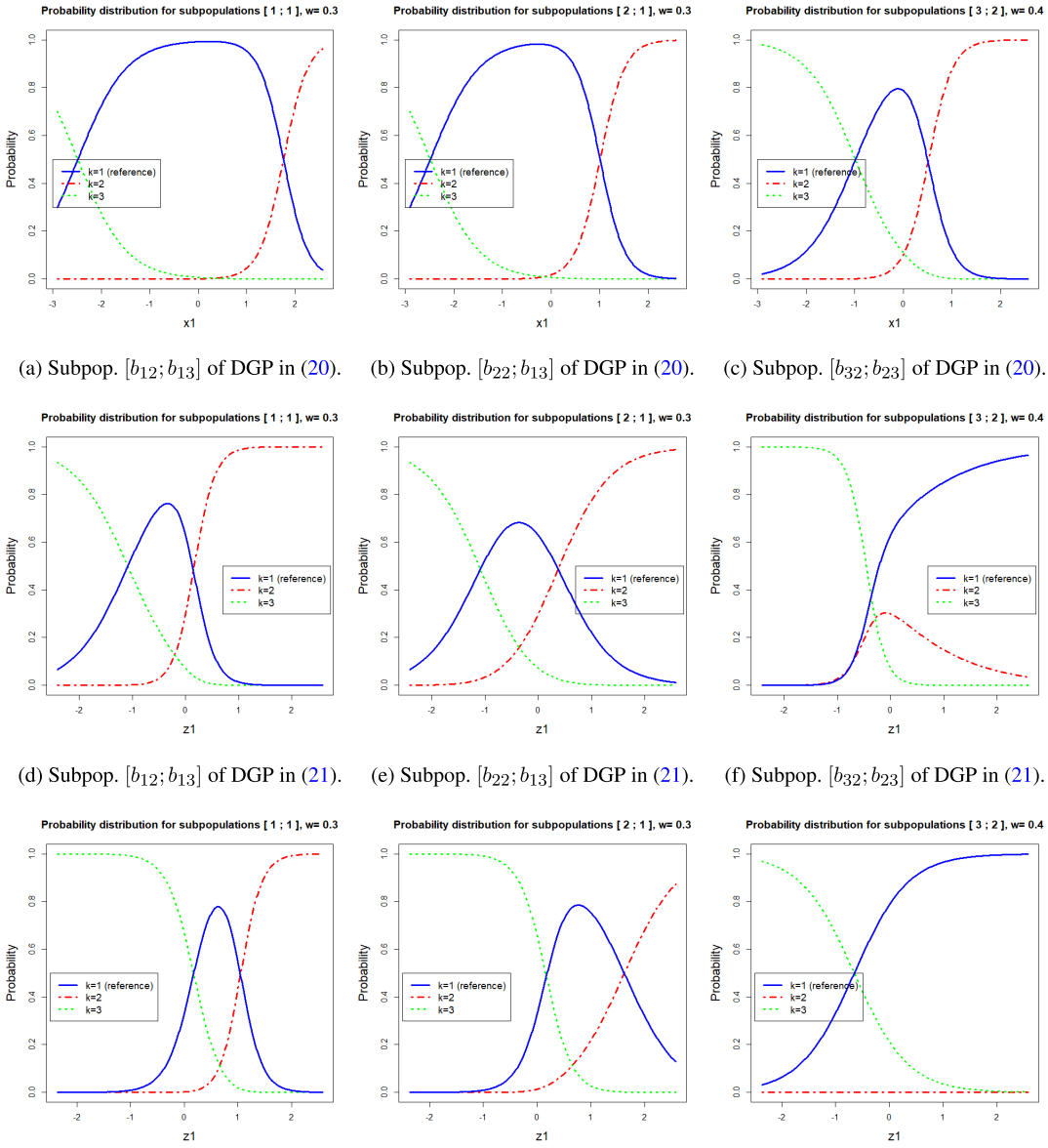
TABLE 2

Fixed and random effects coefficients estimated by MSPEM algorithm in the three different settings. Estimates are reported in terms of mean ± sd, computed on the 500 runs of the simulation study for the fixed effects coefficients and on the runs in which the algorithm identifies $M_2 = 3$ and $M_3 = 2$ (reported in Table 1) for the random effects ones. In order to ease the comparison with the DGPs, true values (TV) of the coefficients used to simulate data are reported under the relative estimates

	$\hat{\alpha}_{1k}$	$\hat{\alpha}_{2k}$	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$
Fixed and random effects coefficients estimated by MSPEM algorithm for the DGP in equation (20)				
$k=2$	$\hat{\alpha}_{12} = 4.096 \pm 0.081$	$\hat{\alpha}_{22} = -3.051 \pm 0.053$	$\hat{b}_{12} = -6.819 \pm 0.182$	$\hat{w}_{12} = 0.300$
			$\hat{b}_{22} = -3.916 \pm 0.109$	$\hat{w}_{22} = 0.300$
			$\hat{b}_{32} = -2.122 \pm 0.099$	$\hat{w}_{32} = 0.400$
	$TV = +4$	$TV = -3$	$TV = (-7, -4, -2)$	$TV = (0.3, 0.3, 0.4)$
$k=3$	$\hat{\alpha}_{13} = -2.067 \pm 0.046$	$\hat{\alpha}_{23} = 2.059 \pm 0.034$	$\hat{b}_{13} = -5.200 \pm 0.089$	$\hat{w}_{13} = 0.599$
			$\hat{b}_{23} = -1.899 \pm 0.048$	$\hat{w}_{23} = 0.401$
	$TV = -2$	$TV = +2$	$TV = (-5, -2)$	$TV = (0.6, 0.4)$
Fixed and random effects coefficients estimated by MSPEM algorithm for the DGP in equation (21)				
$k=2$	$\hat{\alpha}_{12} = -1.195 \pm 0.039$	$\hat{\alpha}_{22} = -2.766 \pm 0.085$	$\hat{b}_{12} = 4.786 \pm 0.121$	$\hat{w}_{12} = 0.300$
			$\hat{b}_{22} = 1.811 \pm 0.071$	$\hat{w}_{22} = 0.301$
			$\hat{b}_{32} = -0.117 \pm 0.134$	$\hat{w}_{32} = 0.399$
	$TV = -1$	$TV = -3$	$TV = (+5, +2, -1)$	$TV = (0.3, 0.3, 0.4)$
$k=3$	$\hat{\alpha}_{13} = -1.672 \pm 0.039$	$\hat{\alpha}_{23} = 1.713 \pm 0.051$	$\hat{b}_{13} = -1.601 \pm 0.057$	$\hat{w}_{13} = 0.600$
			$\hat{b}_{23} = -4.791 \pm 0.210$	$\hat{w}_{23} = 0.400$
	$TV = -2$	$TV = +2$	$TV = (-2, -6)$	$TV = (0.6, 0.4)$
Fixed and random effects coefficients estimated by MSPEM algorithm for the DGP in equation (22)				
$k=2$	$\hat{\alpha}_2 = -5.013 \pm 0.098$	$\hat{b}_{112} = -5.863 \pm 0.236$	$\hat{b}_{212} = 5.091 \pm 0.195$	$\hat{w}_{12} = 0.300$
		$\hat{b}_{122} = -4.700 \pm 0.129$	$\hat{b}_{222} = 2.801 \pm 0.119$	$\hat{w}_{22} = 0.300$
		$\hat{b}_{132} = -8.022 \pm 0.237$	$\hat{b}_{232} = -1.185 \pm 0.079$	$\hat{w}_{32} = 0.400$
	$TV = -5$	$TV = (-6, -4, -8)$	$TV = (+5, +2, -1)$	$TV = (0.3, 0.3, 0.4)$
$k=3$	$\hat{\alpha}_3 = 1.977 \pm 0.040$	$\hat{b}_{113} = 0.739 \pm 0.058$	$\hat{b}_{213} = -3.651 \pm 0.092$	$\hat{w}_{13} = 0.600$
		$\hat{b}_{123} = -0.888 \pm 0.055$	$\hat{b}_{223} = 2.419 \pm 0.160$	$\hat{w}_{23} = 0.400$
	$TV = +2$	$TV = (+1, -1)$	$TV = (-4, +2)$	$TV = (0.6, 0.4)$

to a unique point, and the algorithm will identify a higher number of subpopulations. More details about the impact of the choice of D_k and some insights about how to identify its best choice can be found in Masci, Paganoni and Ieva (2019).

In order to visualize the results, Figure 1 reports the baseline-category logits, computed for each combination of subpopulations across the categories, for the three simulated cases, extracted from one of the 500 runs (randomly chosen). Given the data generating process in equations (20), (21) and (22), the joint distribution of the two random effects coefficients distributions has, in all the three settings, three nonzero weight support points that we express as $[\hat{b}_{m_2 2}; \hat{b}_{m_3 3}]$. In particular, these three support points with their relative weights are $[\hat{b}_{12}; \hat{b}_{13}]$ with weight 0.3, $[\hat{b}_{22}; \hat{b}_{13}]$ with weight 0.3 and $[\hat{b}_{32}; \hat{b}_{23}]$ with weight 0.4. Indeed, there are no groups that, for example, belong to subpopulation 1, regarding $k = 2$, and subpopulation 2, regarding $k = 3$. We report the $2 - D$ visualization of the logits in which on the abscissa we report the covariate $x1$ for the random intercept case and z_1 for random slope and random intercept and slope cases, respectively; we then adjust the baseline-category logits for



(a) Subpop. $[b_{12}; b_{13}]$ of DGP in (20). (b) Subpop. $[b_{22}; b_{13}]$ of DGP in (20). (c) Subpop. $[b_{32}; b_{23}]$ of DGP in (20).

(d) Subpop. $[b_{12}; b_{13}]$ of DGP in (21). (e) Subpop. $[b_{22}; b_{13}]$ of DGP in (21). (f) Subpop. $[b_{32}; b_{23}]$ of DGP in (21).

(g) Subpop. $[\mathbf{b}_{12}; \mathbf{b}_{13}]$ of DGP in (22). (h) Subpop. $[\mathbf{b}_{22}; \mathbf{b}_{13}]$ of DGP in (22). (i) Subpop. $[\mathbf{b}_{32}; \mathbf{b}_{23}]$ of DGP in (22).

FIG. 1. *Baseline-category logits estimated by MSPEM algorithm for the three DGPs in equations* (20), (21) *and* (22), *first, second and third row, respectively. Each row reports the three combinations with nonzero weight of the two random effects distributions, that is, the one relative to category* $k = 2$ *and the one relative to category* $k = 3$. *For the three cases, respectively, panels* (*a*), (*d*) *and* (*g*) *report the logits estimated for subpopulation* 1 *relative to category* $k = 2$ *and subpopulation* 1 *relative to category* $k = 3$; *panels* (*b*), (*e*) *and* (*h*) *report the logits for subpopulation* 2 *relative to category* $k = 2$ *and subpopulation* 1 *relative to category* $k = 3$ *and panels* (*c*), (*f*) *and* (*i*) *report the logits for subpopulation* 3 *relative to category* $k = 2$ *and subpopulation* 2 *relative to category* $k = 3$.

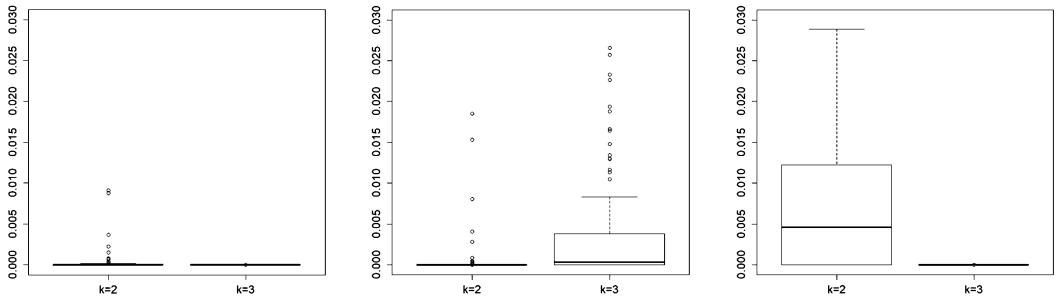
the average effect of the second covariate.[6] We observe that, while from panel (a) to panel (h) of Figure 1, all categories have positive probabilities across all subpopulations; panel (i) represents a case in which the probability that an observation y_{ij} of a group belonging to this

[6]This choice is due to the fact that we are interested in visualizing the trends of the logits for the different values of the random effects coefficients, that is, the intercept and the slope relative to z_1.

subpopulation is equal to $k = 2$, that is, π_{ij2}, is constantly almost null for any value of the covariates.

Lastly, we can evaluate the uncertainty of classification (with which the algorithm classifies groups into subpopulations) by measuring the entropy of the rows of the matrices W_k, for $k = \{2, 3\}$. In the best case, that is, when the algorithm assigns each group i to a subpopulation m_k, relative to category k with probability 1, each row of the matrix W_k would be composed of $M_k - 1$ values equal to 0 and a value equal to 1. In this scenario the entropy $E_i = -\sum_{m_k=1}^{M_k} W_{im_k k} \ln(W_{im_k k})$ of each row i of the matrix W_k would be equal to 0. The more the distribution of the weights is uniform on the M_k mass points, the higher is the entropy and, therefore, the higher is the uncertainty of classification. The worst case happens when the distribution of the weights of a group i is uniform on the M_k subpopulations ($W_{im_k} = 1/M_k$ for $m_k = 1, \ldots, M_k$) which corresponds to an entropy $E_i = -\sum_{m_k=1}^{M_k} 1/M_k \ln(1/M_k) = -\ln(1/M_k)$. Furthermore, the analysis of the entropy of the conditional weights matrices W_k might help to select the value of the tuning parameter D_k, suggesting a lower bound for D_k that minimizes the entropy.[7] In order to ease the interpretation and the comparison across methods, we refer to the normalised entropy, that is, the value of the entropy divided by the maximum possible entropy value ($-\ln(1/M_k)$). The normalised entropy ranges between 0 and 1. Figure 2 reports the distribution of the normalised entropy of W_{i2} and W_{i3}, for $i = 1, \ldots, I$, for the three simulated cases, averaged on the runs in which the algorithm identifies $M_2 = 3$ and $M_3 = 2$.[8]

We observe that the entropy level is always very low (considering that maximum uncertainty corresponds to the maximum entropy of 1), suggesting that, for the simulated data, the MSPEM algorithm classifies groups into subpopulations with a low level of uncertainty (i.e., it clearly distinguishes the presence of patterns within the data). In particular, by comparing the three panels of Figure 2, we note that when the complexity of the random component



(a) Normalised entropy of matrix W for the random intercept case.

(b) Normalised entropy of matrix W for the random slope case.

(c) Normalised entropy of matrix W for the random intercept and slope case.

FIG. 2. *Boxplots of the normalised entropy of W_k, for $k = \{2, 3\}$, measured for each group, obtained by averaging 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*).

[7]Note that this entropy-based method only drives the choice of the minimum value of D_k but not the maximum one. Indeed, by increasing D_k, the mass points of random effects will easily collapse to a low number of final mass points, and the algorithm will assign each group to a subpopulation with a very low level on uncertainty (having it no choice), but this is clearly not an indicator of goodness of the model. On the opposite, if for smaller values of D_k the algorithm is still able to assign groups to the subpopulations with a low level of uncertainty, this means that the groups are well distinguished, even at this deepness.

[8]Entropy values relative to $k = 2$ are normalised by $-\ln(1/3)$ (since $M_k = 3$), while entropy values relative to $k = 3$ are normalised by $-\ln(1/2)$ (since $M_k = 2$).

increases (and this happens accordingly to the order (a)—only random intercept, (b)—only random slope and (c)—random intercept and slope), the entropy also increases. Further analysis show that the entropy computed on the runs in which the algorithm identifies more than $M_2 = 3$ and $M_3 = 2$ subpopulations as well as the entropy obtained with smaller tuning parameters D_k, for $k = \{2, 3\}$, are higher, suggesting that the algorithm does not clearly distinguish the belonging of groups to the subpopulations which result to be too close with respect to the variability within the data. In this sense the entropy of W provides a good indicator both for the choice of the algorithm parameters and for the evaluation of the final model.

In the context of identifying good values for the tuning parameter D_k, a last note regards the maximum likelihood values found by the MSPEM algorithm. In Supplementary Material A in Masci, Ieva and Paganoni (2022), we prove that the proposed EM algorithm increases the likelihood at every step, considering a given number of support points M_k for the random effects distribution. The agglomerated phase, in which we force the support points closest than D_k to collapse, reduces the random effects distribution cardinality and induces the likelihood to decrease. Indeed, every time that two mass points collapse, the parameters space on which we look for the maximum likelihood reduces to a subspace of the previous one. Since the MSPEM algorithm's aim is to cluster groups of observations, we tolerate lower values of the likelihood, but we gain in the identification of a latent distribution of groups. In this sense we expect the maximum likelihood to decrease when D_k increases (and, therefore, M_k decreases). In this perspective the best value of D_k is the one that corresponds to the elbow of the maximum likelihood trend: we look for the minimum number of subpopulations (i.e., maximum clustering) whose parameters still provide high values of the likelihood (i.e., minimum loss in the likelihood).

Results of the presented simulation study show the capability of the MSPEM algorithm to identify the simulated subpopulations of groups and to estimate the model parameters. Nonetheless, Table 2 shows that the bias in some estimates is not completely negligible. In the light of the *asymptotically* unbiased property of the ML estimates in multinomial regression, we make a further simulation study in which we reproduce the one presented in this section but consider 500 observations within each group $i = 1, \ldots, 100$, instead of 200. Supplementary Material C in Masci, Ieva and Paganoni (2022) reports the results and evidences that, by increasing the number of observations, the MSPEM algorithm provides much more unbiased and stable estimates and lower level of uncertainty. In this perspective it is worth to notice that, even if a number of observations equal to 200 constitutes a relatively big group size, the simulated model parameters generate unbalanced categories, some of which present very low absolute frequencies.

4. Case study: University student dropout across engineering degree programmes. We apply the MSPEM algorithm to data about PoliMI students in order to classify different profiles of engineering students and to identify subpopulations of similar degree programmes. We focus on all concluded careers of students enrolled in an engineering programme of PoliMI in the academic year between 2010/2011 and 2015/2016. PoliMI offers 23 different engineering programmes, and students are structurally nested within those programmes. We exclude from the study four degree programmes having few students enrolled—less than 200. 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. We distinguish between two types of dropout:

- *early dropout*—occurs when the student drops within the *third* semester after the enrolment;
- *late dropout*—occurs when the student drops after the *third* semester after the enrolment.

TABLE 3
List and explanation of variables at student level included in the MSPEM model

Variable	Description	Type of variable
`Status`	Status of concluded career	Three-levels 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-levels factor

We make this distinction because we believe the determinants that drive these two types of dropout might be structurally different. The sample contains 16.2% of early dropout students, 16.5% of late dropout students and 67.3% of graduate ones.

The MSPEM algorithm, applied to these data, models the probability of being *early* or *late dropout* student, given student characteristics and early career information, and considers the nested structure of students within the 19 degree programmes. In particular, we are interested in identifying subpopulations of degree programmes, depending on their effects on *early* and *late dropout* probability.

Regarding student characteristics, besides the status of the concluded career and the degree programme the student is enrolled in, we consider the number of European Credit Transfer System credits (ECTS), that is, 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). We consider the information at the first semester of career because it is observable for all students (either dropout or graduate) and guided by the aim of predicting student dropout as soon as possible, that is, at the beginning of the student career. Previous studies on these data reveal that the number of credits obtained at the first semester of career is the most significant covariate for predicting student dropout (Cannistrà et al. (2021), Pellagatti et al. (2021), Fontana et al. (2021)). Table 3 reports the variables considered in the analysis with their explanation.

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

$$Y_{ij} = \begin{cases} \text{Graduate} & \pi_{ij1}, \\ \text{Early dropout} & \pi_{ij2}, \\ \text{Late dropout} & \pi_{ij3}, \end{cases} \tag{23}$$

where

$$\pi_{ijk} = P(Y_{ij} = k) = \frac{\exp(\eta_{ijk})}{1 + \sum_{k=2}^{3} \exp(\eta_{ijk})} \quad \text{for } k = 1, \dots, 3 \tag{24}$$

and

$$\eta_{ijk} = \begin{cases} \mathbf{x}'_{ij}\boldsymbol{\alpha}_k + \delta_{ik} & k = 2, 3, \\ 0 & k = 1. \end{cases} \tag{25}$$

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

TABLE 4
Fixed and random effects coefficients estimated by MSPEM algorithm for student dropout prediction

	$\hat{\alpha}_{1k}$ (Gender)	$\hat{\alpha}_{2k}$ (TotalCredits1.1)	$\hat{b}_{m_k k}$ (random intercept DegProg)	$\hat{w}_{m_k k}$ (weight)
$k=2$	$\hat{\alpha}_{12}=-0.153$	$\hat{\alpha}_{22}=-2.704$	$\hat{b}_{12}=-2.841$	$\hat{w}_{12}=0.482$
			$\hat{b}_{22}=-2.423$	$\hat{w}_{22}=0.272$
			$\hat{b}_{32}=-2.096$	$\hat{w}_{32}=0.193$
			$\hat{b}_{42}=-1.586$	$\hat{w}_{42}=0.053$
$k=3$	$\hat{\alpha}_{13}=-0.685$	$\hat{\alpha}_{23}=-1.899$	$\hat{b}_{13}=-2.152$	$\hat{w}_{13}=0.210$
			$\hat{b}_{23}=-1.733$	$\hat{w}_{23}=0.421$
			$\hat{b}_{33}=-1.219$	$\hat{w}_{33}=0.262$
			$\hat{b}_{43}=-0.880$	$\hat{w}_{43}=0.107$

We run the MSPEM algorithm with `tollR`=`tollF`=10^{-2}, `itmax`=60, `it1`=20, $\tilde{w}=0$ (because we do not want to fix a minimum number of degree programmes within each subpopulation) and $D_k=0.3$, for $k=2,3$. Since $I=19$ is reasonably low, the algorithm starts by considering $M_k^*=19$ starting support points, for $k=\{2,3\}$, identified by I multinomial logistic regression. Starting weights are uniformly distributed on these 19 support points, for $k=\{2,3\}$. The algorithm converges in *seven* iterations and identifies *four* supopulations for both categories $k=2$ (early dropout) and $k=3$ (late dropout). Table 4 reports the estimated model parameters.

By looking at Table 4, we observe that females have, on average, lower probability of both early and, especially, late dropout than males (-0.153 and -0.685, respectively). The number of credits obtained at the first semester is inversely proportional to the probability of both early and late dropout: the higher is the value of `TotalCredits1.1`, the lower is the estimated probability of late and, especially, early dropout (-1.899 and -2.704, respectively). Regarding the random intercepts, Table 4 reports the random intercepts associated to the four subpopulations, for each k, with their weights, ordered increasingly. The distributions of the 19 degree programmes across the identified subpopulations relative to $k=2,3$ are reported in Table 5. Each degree program belongs to the subpopulation for which the posterior conditional weight is the highest. For each k, subpopulation 1 contains the degree programmes associated to the lowest random intercept, that is, degree programmes in which students are less likely to dropout with respect to the average. On the opposite, subpopulation 4 contains the degree programmes associated to the highest random intercept, that is, degree programmes in which students are more likely to dropout with respect to the average.

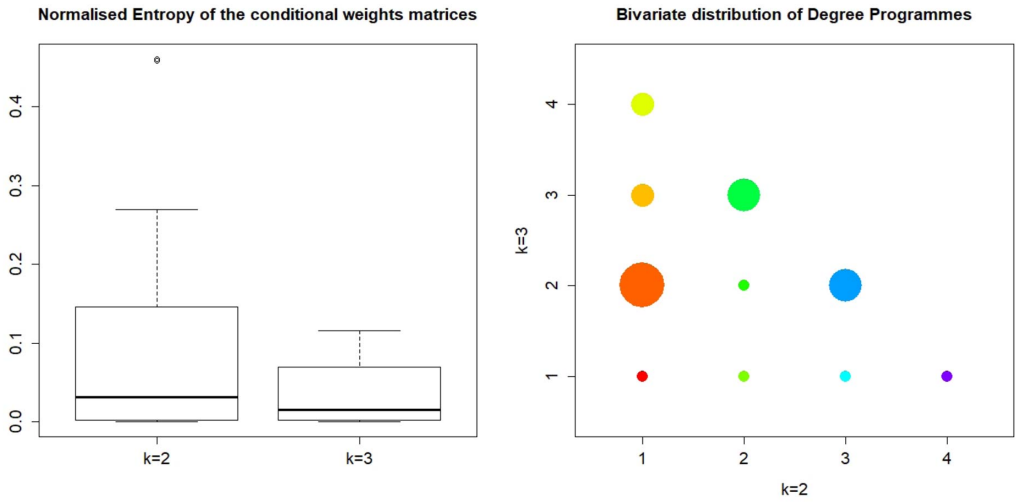
Regarding early dropout (i.e., $k=2$), from Table 5 we observe that the most numerous subpopulation (subpopulation 1, containing eight degree programmes out of $19-\hat{w}_{12}=0.482$) is the one associated to the lowest early dropout probability, while the other three subpopulations contain degree programmes in which students are more likely to early drop out, net to their personal characteristics. In particular, biomedical engineering is identified as an outlier, associated to the highest early dropout probability.[9] For late dropout (i.e., $k=3$), the most numerous subpopulation is subpopulation 2 (containing eight degree programmes out of $19-\hat{w}_{23}=0.421$), and, with respect to it, there is a subpopulation of degree programmes

[9]This result in reasonable and expected since in Italy many students who can not access the medicine faculty, given to the selective entrance exam, attend different faculties, for example, biomedical engineering, waiting to be admitted to medicine.

TABLE 5

Distribution of the 19 *degree programmes across the four identified subpopultions relative to* $k = 2, 3$. *For each* k *the order of the four subpopulations is coherent to the one of the estimated random intercepts in Table* 4

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		



(a) Normalised entropy of the conditional weights matrices W_k, for $k = \{2, 3\}$.

(b) Degree programmes distribution across subpopulations.

FIG. 3. *Panel (a) reports the boxplots of the normalised entropy of W_k, for $k = \{2, 3\}$, measured for each degree course*; *Panel (b) reports the distribution of degree programmes across the subpopulations relative to $k = \{2, 3\}$ (each degree course belongs to a subpopulation relative to $k = 2$ and to an other one relative to $k = 3$). Bubble size is proportional to the number of degree programmes belonging to the couple $(m_{\phi 1}, m_{\psi 2})$, for $\phi, \psi = 1, \ldots, 4$.*

associated to a lower late dropout probability and two subpopulations associated to higher late dropout probability. In particular, electronic engineering and engineering of computing system (subpopulation 4) are the ones in which students are more likely to late drop out.

In order to evaluate the uncertainty of classification of the MSPEM algorithm in this case study, panel (a) in Figure 3 reports the normalised entropy distribution of the weight matrices W_k, for $k = 2, 3$, computed as the entropy divided by the maximum entropy relative to the case of four subpopulations, that is, $-\log(1/4) = 1.38$. The normalised entropies of W_1 and, especially, W_2 are low (in particular, normalised entropy median and mean are 0.031 and 0.109 for $k = 2$ and 0.015 and 0.035 for $k = 3$, respectively), suggesting that most of degree programmes are associated to a subpopulation with a very low level of uncertainty. This result also drove our choice of $D = 0.3$, since it is a threshold that allows to distinguish the highest number of subpopulations with a low level of uncertainty.[10] Lastly, panel (b) in Figure 3 gives us a graphical representation of the correlation among the subpopulations distributions relative to $k = 2, 3$. For each couple of mass points $(m_{\phi 1}, m_{\psi 2})$, for $\phi, \psi = 1, \ldots, 4$, bubble size is proportional to the number of degree programmes that belong to this couple. It does not emerge a clear correlation between the two distributions (considering that most of the observations are in the first two subpopulations for both $k = 2, 3$), but we notice that there are no degree programmes associated to both high early and late dropout probability (i.e., there are no subpopulations belonging to couples (m_{32}, m_{33}), (m_{42}, m_{33}), (m_{32}, m_{43}) and (m_{42}, m_{43})), suggesting that degree programmes in which students are more likely to early drop out are also less likely to late drop out and vice-versa.

4.1. *Comparison with MCMCglmm method.* As we anticipated in the Introduction, Hadfield et al. (2010) propose an MCMC method for multiresponse generalized linear mixed-models that provides a robust strategy for marginalizing the random effects. Here, we apply

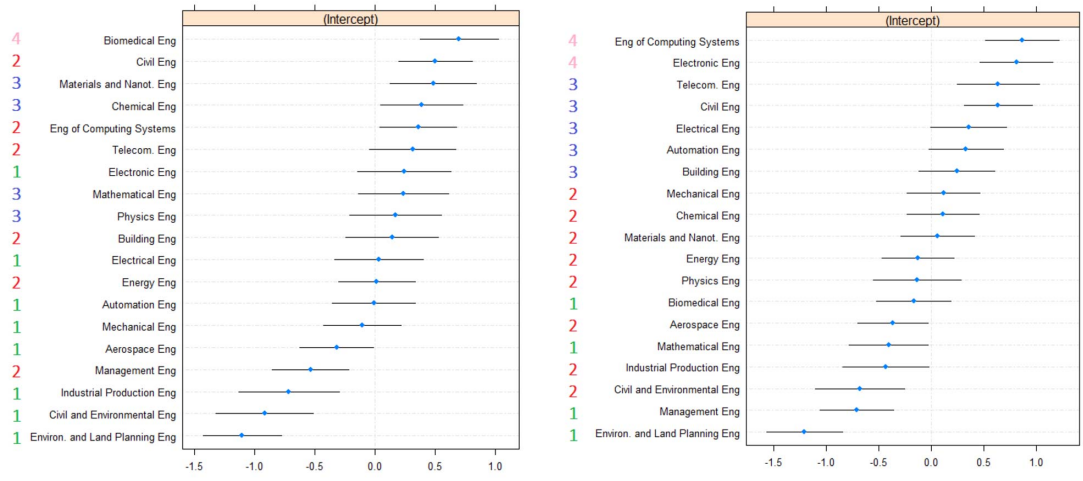
[10] Note that for small variations of D around 0.3, for example, ± 0.05, results remain consistent.

TABLE 6
Fixed effects estimates of the MCMCglmm *method*

	Variable name	post.mean	$l-95\%$ CI	$u-95\%$ CI	pMCMC
$k=2$	`Intercept`	−2.552	−2.854	−2.269	<0.001**
	`Gender`	−0.027	−0.106	0.153	0.769
	`TotalCredits1.1`	−2.797	−2.884	−2.702	<0.001**
$k=3$	`Intercept`	−2.354	−2.672	−2.049	<0.001**
	`Gender`	−0.634	−0.464	−0.802	<0.001**
	`TotalCredits1.1`	−2.135	−2.198	−2.067	<0.001**

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

the relative *MCMCglmm* R function to the PoliMI case study in order to compare the results with the ones obtained with the MSPEM algorithm. We run the *MCMCglmm* function with the same set of variables and assumptions selected for the MSPEM algorithm, without specifying any prior, with 30,000 MCMC iterations and a burnin of 2000. Fixed effects estimates are reported in Table 6, while random intercepts with their confidence intervals are shown in Figure 4. Clearly, since the two methods assume the random effects coefficients to follow different distributions, we obtain two different types of results: the MSPEM algorithm identifies a latent structure at the degree programmes level, clustering degree programmes into subpopulations; the MCMCglmm method estimates a single intercept for each degree programme, obtaining a ranking of degree courses. By looking at Table 6, we observe that the estimated coefficients relative to `Gender` and `TotalCredits1.1`, for both $k=2,3$, are coherent with the ones obtained by the MSPEM algorithm, shown in Table 4. This result is in line with the stability theory about fixed effects coefficients that result not to be affected by random effects distributions. Regarding the estimated random intercepts, we are interested in seeing whether the subpopulations, identified by the MSPEM algorithm, are coherent with the ranking of the MCMCglmm intercepts. Results are satisfyingly consistent. For early dropout, Biomedical Engineering, which composes the "outlier" Subpopulation 4,



(a) Degree programmes intercepts for Early Dropout.

(b) Degree programmes intercepts for Late Dropout.

FIG. 4. *Panels* (*a*) *and* (*b*) *report the* MCMCglmm *estimated intercepts with their confidence intervals relative to the* 19 *degree programmes for* $k=2$ (*Early Dropout*) *and* $k=3$ (*Late Dropout*), *respectively. To ease the comparison with MSPEM results, degree programmes on the vertical axes in both panels are sided by their associated subpopulation in Table* 5. *Colours are only intended to help in the visualization.*

is also the first in the intercepts ranking in panel (a) of Figure 4, resulting to be the degree programme associated to the highest early dropout probability by both the methods. Equivalently, Subpopulation 1, associated to the lowest early dropout probability, contains degree programmes that are at the bottom of the ranking in panel (a) of Figure 4, except for Electronic and Electrical Engineering. For late dropout the consistency of results between the two methods is even sharper: the four subpopulations identified by the MSPEM algorithm clusters the 19 degree programmes according to the ranking shown in panel (b) of Figure 4, identifying the first two degree programmes, that is, Engineering of Computing Systems and Electronic Engineering, as components of the subpopulation associated to the highest late dropout probability. To ease the comparison between the two methods, degree programmes on the vertical axes in Figure 4 are sided by their associated subpopulation in Table 5.

The identification of subpopulations might also be interpreted as a robustness check tool for the groups' ranking that we obtain when assuming normal distributed random effects: in the fully parametric context, we have no evidence to document differences and, consequently, to create a statistically significant ranking between groups whose associated confidence intervals are overlapped. Equivalently, we have no evidence to identify significant differences between those groups whose confidence intervals contain zero and the average. To this perspective, groups that have confidence intervals clearly far from zero or from the ones of other groups are expected to belong to "outlier" subpopulations, while groups that have confidence intervals overlapped to many other ones are expected to be misclassified within subpopulations.

Lastly, regarding possible implications, the identified subpopulations could be characterized in terms of degree programmes information in order to identify potential drivers of the different students dropout rates.

Comparison of MSPEM and MCMCglmm goodness of fit. From an interpretative point of view, the *MSPEM* and the *MCMCglmm* algorithms produce comparable results and identify coherent associations between the fixed effects coefficients and the response and similar patterns in the degree programmes effects. In order to evaluate and compare the goodness of fit of the two methods, we compute their relative misclassification tables.[11]

The two methods reveal similar predictive performances. Error rates are 21.6% for *MCMCglmm* and 23.3% for *MSPEM*, respectively. The *MSPEM* algorithm that considers, for each contrast, the four identified subpopulations as representative of the different dynamics across the 19 degree programmes, fits the data comparably well to the *MCMCglmm* that considers a different dynamic for each single degree programme.

5. Concluding remarks and future perspectives. In this paper we propose a semiparametric multinomial mixed-effects linear model, together with an expectation-maximization

TABLE 7
Misclassification tables relative to MSPEM (*left tabular*) *and* MCMCglmm (*right tabular*) *predictions*

	obs D1	obs D2	obs G
pred D1	0.095	0.063	0.019
pred D2	0.035	0.038	0.017
pred G	0.032	0.066	0.635

	obs D1	obs D2	obs G
pred D1	0.100	0.058	0.018
pred D2	0.032	0.047	0.017
pred G	0.030	0.061	0.637

[11] In the perspective of evaluating the goodness of fit, we report in Table 7 the misclassification tables computed on the same set of data we used to train the models.

algorithm to estimate its parameters, to model university student dropout. We assume the random effects of the mixed-effects model to follow a discrete distribution with an unknown number of support points. Considering a multinomial response variable assuming K categories, the model is identified by $K-1$ p-dimensional vectors of fixed effects coefficients (where p is the number of fixed effects covariates) and $K-1$ q-variate random effects distributions (where q is the number of random effects covariates) with $M_{k'}$ support points, for $k'=1,\ldots,K-1$. This modelling allows us to identify a latent structure at the highest level of the hierarchy in which groups collapse into a finite and a priori unknown number of subpopulations. In particular, we estimate a subpopulations distribution related to each of the $K-1$ baseline-category logits. Moreover, in a multinomial response context in which classical gaussian random effects are analytically and numerically difficult to be integrated out, our proposed discrete random effects allow us to express the marginal distribution of the response as a weighted sum, avoiding difficult integration problems.

We develop the MSPEM algorithm guided by the aim of profiling university students and of identifying subpopulations of engineering degree programmes, standing on their effects on their students dropout probability. In particular, we model different profiles of engineering university students, considering their early career information and their nested structure within degree programmes. The algorithm identifies subpopulations of degree programmes in which students are more/less likely to early or late drop their studies. We compare our results with the ones obtained by applying a fully parametric method, the *MCMCglmm*, to the same dataset, underlining similarities and differences and exploiting the different types of results provided by fully parametric and semiparametric methods.

From an interpretative point of view, the MSPEM algorithm can be seen as an in-built clustering tool, and the subpopulations identified by it represent an alternative to the ranking provided by classical parametric mixed-effects models. In the general context of educational data, lower education students are nested within classes and schools, whose cardinality is often very high. In this perspective, identifying subpopulations of classes/schools, instead of a ranking of hundreds or thousands of observations whose estimates are sometimes so closed to be indistinguishable, might be easier and more effective. Moreover, the number of subpopulations is not fixed a priori, but it is estimated by the algorithm based on the threshold value of the euclidean distance D between mass points. This allows us to build the subpopulations standing on how much we want to be sensitive to the differences at the highest level of the hierarchy, depending on the aims of the study.

This work enters in the literature about mixed-effects models with discrete random effects (Aitkin (1999), Hartzel (2000), Masci, Paganoni and Ieva (2019)) and proposes a novel method that deals with multinomial responses. The MSPEM algorithm has been developed for our specific case study, but, sharpening its settings and augmenting its complexity, it is possible to refine this study adding, for example, further students or degree programmes information among the fixed or random effects covariates. More generally, it can be applied to any classification problem dealing with a multiple categories response and hierarchical data, a context in which the literature is still poor and quite challenging.

The development of mixed-effects multinomial models with discrete random effects is still at its beginning. Several issues remain unresolved, and further developments are needed regarding the random effects structure assumptions, the coefficients significance and the variability and stability of estimates. First and foremost, at the current state of the art, random effects are considered independent across categories. Since the random effects from different logits arise from the same subjects, this assumption may be unrealistic and too restrictive. Especially, panel (b) in Figure 3 shows that, in the case study, the distributions of degree programmes on the random effects support points relative to the two contrasts are not independent. This evident pattern in the bivariate distribution reveals that we are not taking into

account an important part of the data structure. Therefore, a first future perspective regards the possibility to relax the independence assumption across categories, extending the proposed method to deal with various complex dependence structures for modelling more realistic correlations within real data. A second aspect that deserves attention regards the measurement of the estimates' variability, that is, the estimation of fixed and random effects coefficients standard errors and of their significance. At the current state of the art, the MSPEM algorithm does not provide estimates of standard errors. Further work is needed to assess the significance of fixed effects coefficients by means, for example, of likelihood-ratio tests and to evaluate the relevance of the identified random effects structure. To this end, in parametric multilevel models the variance partition coefficient measures the percentage variability explained by the data highest level structure. Its extension to semiparametric multilevel models have also been proposed (Goldstein, Browne and Rasbash (2002), Rights and Sterba (2016), Masci et al. (2021)), and we are working to adapt it to the case of multinomial semiparametric mixed-effects models.

SUPPLEMENTARY MATERIAL

Supplement to "Semiparametric multinomial mixed-effects models: A university students profiling tool" (DOI: 10.1214/21-AOAS1559SUPP; .pdf). Supplementary Material to the paper is provided in Masci, Ieva and Paganoni (2022) and contains the following three sections. *SM-A: Proof of the increasing likelihood property of the MSPEM algorithm*. In this section, we show the proof of the increasing likelihood property of the MSPEM algorithm. In particular, we show how we derive equations (11) and (13) and why these two equations constitute the steps of the Expectation-Maximization algorithm. *SM-B: Technical details about the MSPEM algorithm*. Technical details about the MSPEM algorithm are given. We give insights about the discrete distribution support points initialization, the support points collapse criteria and the algorithm convergence criteria. *SM-C: Simulation results under different group sizes*. In this section, we reproduce the simulation study shown in Section 3, but considering 500 observations within each group $i = 1, \ldots, 100$, instead of 200. Since ML estimates in multinomial regression are only *asymptotically* unbiased, we implement this further simulation in order to evaluate the improvement in the estimates given to the increased number of observations.

REFERENCES

AGRESTI, A. (2018). *An Introduction to Categorical Data Analysis*. Wiley, New York.

AINA, C. (2013). Parental background and university dropout in Italy. *High. Educ.* **65** 437–456.

AITKIN, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55** 117–128. MR1705676 https://doi.org/10.1111/j.0006-341X.1999.00117.x

ALJOHANI, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. *High. Educ. Stud.* **6** 1–18.

ANDERSON, D. A. and AITKIN, M. (1985). Variance component models with binary response: Interviewer variability. *J. Roy. Statist. Soc. Ser. B* **47** 203–210. MR0816084

ANDERSON, C. J., KIM, J.-S. and KELLER, B. (2013). Multilevel modeling of categorical response variables. In *Handbook of International Large-Scale Assessment*: *Background*, *Technical Issues*, *and Methods of Data Analysis* 481–519.

ANVUR (2018). Rapporto biennale sullo stato del sistema universitario e della ricerca. Available at https://www.anvur.it/rapporto-biennale/rapporto-biennale-2018.

BARBU, M., VILANOVA, R., VICARIO, J., PEREIRA, M. J., ALVES, P., PODPORA, M., KAWALA-JANIK, A., PRADA, M., DOMINGUEZ, M. et al. (2019). Data mining tool for academic data exploitation: Publication report on engineering students profiles. ERASMUS+ KA2/KA203.

BELLOC, F., MARUOTTI, A. and PETRELLA, L. (2011). How individual characteristics affect university students drop-out: A semiparametric mixed-effects model for an Italian case study. *J. Appl. Stat.* **38** 2225–2239. MR2843254 https://doi.org/10.1080/02664763.2010.545373

BOCK, R. D. and AITKIN, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46** 443–459. MR0668311 https://doi.org/10.1007/BF02293801

BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo em algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 265–285.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

BRESLOW, N. E. and LIN, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82** 81–91. MR1332840 https://doi.org/10.1093/biomet/82.1.81

CANNISTRÀ, M., MASCI, C., IEVA, F., AGASISTI, T. and PAGANONI, A. M. (2021). Early-predicting dropout of university students: an application of innovative machine learning and multilevel statistical techniques *Studies in Higher Education* in press.

COULL, B. A. and AGRESTI, A. (2000). Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics* **56** 73–80. https://doi.org/10.1111/j.0006-341x.2000.00073.x

DE FREITAS, S., GIBSON, D., DU PLESSIS, C., HALLORAN, P., WILLIAMS, E., AMBROSE, M., DUNWELL, I. and ARNAB, S. (2015). Foundations of dynamic learning analytics: Using university student data to increase retention. *Br. J. Educ. Technol.* **46** 1175–1188.

DE LEEUW, J., MEIJER, E. and GOLDSTEIN, H. (2008). *Handbook of Multilevel Analysis*. Springer, Berlin.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

DOS SANTOS, D. M. and BERRIDGE, D. M. (2000). A continuation ratio random effects model for repeated ordinal responses. *Stat. Med.* **19** 3377–3388.

FONTANA, L., MASCI, C., IEVA, F. and PAGANONI, A. (2021). Performing learning analytics via generalized mixed-effects trees *Data* **6** 7–74.

GOLDSTEIN, H. (2011). *Multilevel Statistical Models* **922**. Wiley, New York.

GOLDSTEIN, H., BROWNE, W. and RASBASH, J. (2002). Partitioning variation in multilevel models. *Underst. Stat.* **1** 223–231.

GOLDSTEIN, H. and RASBASH, J. (1996). Improved approximations for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser. A* **159** 505–513. MR1413664 https://doi.org/10.2307/2983328

HADFIELD, J. D. et al. (2010). Mcmc methods for multi-response generalized linear mixed models: The mcmcglmm R package. *J. Stat. Softw.* **33** 1–22.

HARTZEL, J. S. (2000). Random effects models for nominal and ordinal data.

HARTZEL, J., AGRESTI, A. and CAFFO, B. (2001). Multinomial logit random effects models. *Stat. Model.* **1** 81–102.

HEINEN, T. (1996). *Latent Class and Discrete Latent Trait Models*: *Similarities and Differences*. Sage, Thousand Oaks.

LINDSAY, B. G. (1983a). The geometry of mixture likelihoods: A general theory. *Ann. Statist.* **11** 86–94. MR0684866 https://doi.org/10.1214/aos/1176346059

LINDSAY, B. G. (1983b). The geometry of mixture likelihoods. II. The exponential family. *Ann. Statist.* **11** 783–792. MR0707929 https://doi.org/10.1214/aos/1176346245

MASCI, C., IEVA, F. and PAGANONI, A. M. (2022). Supplement to "Semiparametric multinomial mixed-effects models: A university students profiling tool." https://doi.org/10.1214/21-AOAS1559SUPP

MASCI, C., PAGANONI, A. M. and IEVA, F. (2019). Semiparametric mixed effects models for unsupervised classification of Italian schools. *J. Roy. Statist. Soc. Ser. A* **182** 1313–1342. MR4027363 https://doi.org/10.1111/rssa.12449

MASCI, C., IEVA, F., AGASISTI, T. and PAGANONI, A. M. (2021). Evaluating class and school effects on the joint student achievements in different subjects: A bivariate semiparametric model with random coefficients. *Comput. Statist.* 1–41. https://doi.org/10.1007/s00180-021-01107-1

MCCULLOCH, C. E. (1994). Maximum likelihood variance components estimation for binary data. *J. Amer. Statist. Assoc.* **89** 330–335.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170. MR1436105 https://doi.org/10.2307/2291460

MCCULLOCH, C. E. and SEARLE, S. R. (2001). *Generalized*, *Linear*, *and Mixed Models*. *Wiley Series in Probability and Statistics*: *Texts*, *References*, *and Pocketbooks Section*. Wiley-Interscience, New York. MR1884506

MCCULLOCH, C., LIN, H., SLATE, E. and TURNBULL, B. (2002). Discovering subpopulation structure with latent class mixed models. *Stat. Med.* **21** 417–429.

MUTHÉN, B. (2004). Latent variable analysis. *Sage Handb. Quant. Methodol. Soc. Sci.* **345** 106–109.

NAGIN, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychol. Methods* **4** 139.

NAGIN, D. S., JONES, B. L., LIMA PASSOS, V. and TREMBLAY, R. E. (2018). Group-based multi-trajectory modeling. *Stat. Methods Med. Res.* **27** 2015–2023. MR3807918 https://doi.org/10.1177/0962280216673085

PELLAGATTI, M., MASCI, C., IEVA, F. and PAGANONI, A. M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Stat. Anal. Data Min.* **14** 241–257. MR4303069 https://doi.org/10.1002/sam.11505

PINHEIRO, J. and BATES, D. (2006). *Mixed-Effects Models in S and S-PLUS*. Springer, Berlin.

RAUDENBUSH, S. W. (2004). *HLM* 6: *Hierarchical Linear and Nonlinear Modeling*. Scientific Software International.

RAUDENBUSH, S. W., YANG, M.-L. and YOSEF, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *J. Comput. Graph. Statist.* **9** 141–157. MR1826278 https://doi.org/10.2307/1390617

RIGHTS, J. D. and STERBA, S. K. (2016). The relationship between multilevel models and non-parametric multilevel mixture models: Discrete approximation of intraclass correlation, random coefficient distributions, and residual heteroscedasticity. *Br. J. Math. Stat. Psychol.* **69** 316–343.

RODRÍGUEZ, G. and GOLDMAN, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser. A* **158** 73–89.

SHAW, D. S., LACOURSE, E. and NAGIN, D. S. (2005). Developmental trajectories of conduct problems and hyperactivity from ages 2 to 10. *J. Child Psychol. Psychiatry* **46** 931–942.

SHAW, D. S., GILLIOM, M., INGOLDSBY, E. M. and NAGIN, D. S. (2003). Trajectories leading to school-age conduct problems. *Dev. Psychol.* **39** 189–200. https://doi.org/10.1037//0012-1649.39.2.189

SKRONDAL, A. and RABE-HESKETH, S. (2004). *Generalized Latent Variable Modeling*: *Multilevel*, *Longitudinal*, *and Structural Equation Models*. *Interdisciplinary Statistics*. CRC Press/CRC, Boca Raton, FL. MR2059021 https://doi.org/10.1201/9780203489437

SPIEGELHALTER, D., THOMAS, A., BEST, N. and LUNN, D. (2003). Winbugs user manual.

STEELE, F., STEELE, F., KALLIS, C., GOLDSTEIN, H. and 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. and SECREST, D. (1966). *Gaussian Quadrature Formulas*. Prentice-Hall, Inc., Englewood Cliffs, NJ. MR0202312

R CORE TEAM (2019). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

WOLFINGER, R. and O'CONNELL, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *J. Stat. Comput. Simul.* **48** 233–243.

ZHAO, Y., STAUDENMAYER, J., COULL, B. A. and WAND, M. P. (2006). General design Bayesian generalized linear mixed models. *Statist. Sci.* **21** 35–51. MR2275966 https://doi.org/10.1214/088342306000000015