MOX-Report No. 44/2020

EM algorithm for semiparametric multinomial mixed-effects models

Masci, C.; Ieva, F.; Paganoni A.M.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

EM ALGORITHM FOR SEMIPARAMETRIC MULTINOMIAL MIXED-EFFECTS MODELS

BY CHIARA MASCI[1], FRANCESCA IEVA[1,*] AND ANNA MARIA PAGANONI[1,†]

[1]*MOX - Department of Mathematics, Politecnico di Milano, 20133 Milan (Italy)* *chiara.masci@polimi.it;*
[*]*francesca.ieva@polimi.it;* [†]*anna.paganoni@polimi.it*

This paper proposes an EM algorithm for semiparametric mixed-effects models dealing with a multinomial response. In multinomial mixed-effects models, in order to obtain the marginal distribution of the response, random effects need to be integrated out. In a full parametric context, where random effects follow a multivariate normal distribution, this is often computationally infeasible. We propose an alternative novel semiparametric approach in which random effects follow a multivariate discrete distribution with an a priori unknown number of support points, that is allowed to differ across 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 typical of the integration and, second, to identify a latent structure at the highest level of the hierarchy, where groups are clustered into subpopulations. The paper shows a simulation study to evaluate the method's performance and applies the proposed algorithm to a real case study for predicting higher education student dropout, comparing the results with the ones of a full parametric method.

1. Introduction. Many studies deal with hierarchical data, i.e. 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. McCulloch (1994, 1997); Booth and Hobert (1999)) or approximation methods such us 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),

Keywords and phrases: Semiparametric statistics, Multinomial mixed-effects regression, Unsupervised clustering, Higher education

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 et al., 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 et al., 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 fist- or second-order Taylor expansions (Goldstein and Rasbash, 1996), on a combination of a fully multivariate Taylor expansion and a Laplace approximation (Raudenbush et al., 2000), or using Gauss-Hermite quadrature (Stroud 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 et al., 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), 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 multi-response generalized linear mixed models, to provide a robust strategy for marginalizing the random effects (Zhao et al., 2006). This model is developed in a Bayesian 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, 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 et al. (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 and latent class models (see McCulloch et al. (2002); Muthén (2004); Nagin (1999); Heinen (1996) for discussion).

The advantage introduced by the proposed modelling is twofold: (i) the former is that, by assuming a discrete distribution at the higher 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 higher level of the hierarchy, i.e. the presence of subpopulations among groups. Moreover, this modelling allows to investigate how the latent structure at the higher 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 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 et al. (2001), where the authors generalizes Aitkin (1999) for an ordinal random effects model, treating the random effects in a non-parametric manner, assuming them to follow a discrete distribution. Although there are many parallels with this work, the authors in Hartzel et al. (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.

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, a 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 show the case study and the comparison with the parametric MCMCglmm algorithm; in Section 5 we draw our conclusions and discuss some future perspectives.

2. Methodology. Let consider a multinomial logistic regression model for nested data with a two-level hierarchy (Agresti, 2018; De Leeuw et al., 2008), where each observation j, for $j = 1, \dots, n_i$, 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 $i-$th group. The multinomial distribution with K categories relative to Y_{ij} is the following:

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

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 $Y_{ij} = k$, i.e. π_{ijk}, is given by

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

where the linear predictor $\eta_{ijk} = \mathbf{x}'_{ij}\boldsymbol{\alpha}_k + \mathbf{z}'_{ij}\boldsymbol{\delta}_{ik}$ 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[1], that is $k = 1$. Consequently, each category

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

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 Eq. (2), the log-odds of each response with respect to the reference category are:

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

Logit models for nominal response basically pair each category with a baseline category. We therefore observe $K-1$ contrasts, where each contrast $k', k' = 1, \ldots, K-1$, is characterized by the set of *contrast-specific* parameters $(\boldsymbol{\alpha}_{k'}; \boldsymbol{\delta}_{ik'}$, for $i = 1, \ldots, I)$, that models the probability of Y_{ij} being equal to $k \equiv k' + 1$ with respect to the probability of Y_{ij} being equal to 1 (the reference category)[2]. 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 $\mathbf{A} = (\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K)$ and $\boldsymbol{\Delta}_i = (\boldsymbol{\delta}_{i2}, \ldots, \boldsymbol{\delta}_{iK})$, the conditional distribution of Y_{ij} takes the following form:

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

Assuming that $\mathbf{Y}_i$ and $\mathbf{Y}_i'$ are independent for $i \neq i'$, 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 et al., 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 et al., 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,

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

$\log L = \sum_{i=1}^{I} \log h(\mathbf{Y}_i)$, 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 et al. (2019), we move to a semiparametric framework assuming the coefficients of the random effects to follow a discrete distribution with an a priori unknown number of support points.

2.1. *EM algorithm for semiparametric mixed-effects model for a multinomial response.* Considering a semiparametric framework in which the random effects coefficients follow a discrete distribution with an a priori unknown number of support points, the multinomial logit takes the form:

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

where M_k is the total number of support points of the discrete distribution of $\mathbf{b}$ relative to the $k-$th 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 $w_{m_k} \geq 0$:

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

The discrete distribution P_k, for $k = 2, \ldots, K$, 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 (1983); Lindsay et al. (1983),who proved the existence, discreteness and uniqueness of the semiparametric maximum likelihood estimator of a mixing distribution, in the case of exponential family densities. In particular, Lindsay (1983); Lindsay et al. (1983) 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.

Given this formulation, we propose an Expectation-Maximization 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 $K = 2$, i.e. 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 $\mathbf{Y}$ would take the form:

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

In the case of a generic $K > 2$, assuming the $K-1$ discrete distributions of random effects to be independent, the likelihood in Eq. (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:

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

where w_m is the weight relative to the $m-$th combination of the $(K-1)$ weights relative to the $(K-1)$ contrasts and, analogously, $\mathbf{B}_m$ is the $m-$th 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:

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

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 Eq. (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:

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

where

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

and

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

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

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

where

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

The weight $w_{m_k k}^{(up)}$ is the mean over the I groups of their weights relative to the m_k-th subpopulation, relative to category k. Coefficient $W_{im_k k}$ 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 Eq. (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, i.e.

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

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

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

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

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 $W_{im_k k}$ over m_k, i.e.

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

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

Appendix A reports the increasing likelihood property proof of this parameters update procedure.

During the iterations of the EM algorithm, the reduction of the support points of the random effects discrete distributions is performed. Appendix B reports some 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 behaviour under various random effects distributions and subpopulations sizes (Hartzel et al., 2001). Regarding identifiability issues, a mixture is identifiable if it is uniquely characterized , i.e. 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 3 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; (iii) and one considering both random intercept and random slope. We consider $I = 100$ groups of data, where each group contains 200 observations[3] and we induce the presence of three subpopulations regarding category $k = 2$, i.e. $M_2 = 3$, and two subpopulations regarding category $k = 3$, i.e. $M_3 = 2$. In particular, for $j = 1, \ldots, 200$ and $i = 1, \ldots, 100$, we consider the model

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

where the linear predictor $\boldsymbol{\eta}_{ik} = (\eta_{i1k}, \ldots, \eta_{i200k})$ is generated in the following ways[4]:

[3] 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.

[4] 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.

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

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

$$\boldsymbol{\eta}_{i3} = \begin{cases} -2\mathbf{x}_{1i} + 2\mathbf{x}_{2i} - 5 & i = 1, \ldots, 60 \\ -2\mathbf{x}_{1i} + 2\mathbf{x}_{2i} - 2 & i = 61, \ldots, 100 \end{cases} \tag{20}$$

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

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

$$\boldsymbol{\eta}_{i3} = \begin{cases} -2 + 2\mathbf{x}_{1i} - 2\mathbf{z}_{1i} & i = 1, \ldots, 60 \\ -2 + 2\mathbf{x}_{1i} - 6\mathbf{z}_{1i} & i = 61, \ldots, 100 \end{cases} \tag{21}$$

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

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

$$\boldsymbol{\eta}_{i3} = \begin{cases} +2\mathbf{x}_{1i} + 1 - 4\mathbf{z}_{1i} & i = 1, \ldots, 60 \\ +2\mathbf{x}_{1i} - 1 + 2\mathbf{z}_{1i} & i = 61, \ldots, 100 \end{cases} \tag{22}$$

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 100 runs of the MSPEM algorithm for each of the three settings shown in (20), (21) and (22). We fix $D_k = 1$ as threshold value for the support points collapse criterium , for $k = \{2, 3\}$, `tolR` $=$ `tolF` $= 0.01$, `itmax` $= 50$ and `it1` $= 30$ (see Appendix B). In all the runs, the algorithm converges in a number of iterations that ranges between 5 and 10. Table 1 reports the number of runs out of 100 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.

TABLE 1

MSPEM algorithm performance across the 100 runs for each of the three cases. First column reports the number of runs out of 100 in which the algorithm identifies the correct number of subpopulations that are simulated in the data generating process (DGP) in Eq.s (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	94 out of 100	91 out of 94
(ii) Random slope case	91 out of 100	85 out of 91
(iii) Random intercept and slope case	84 out of 100	60 out of 84

The algorithm correctly identifies the simulated number of subpopulations and classifies groups into these subpopulations in more than 85% of the runs, for all 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 3 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 (that is 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 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 et al. (2019).

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 100 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.

	$\hat{\alpha}_{1k}$	$\hat{\alpha}_{2k}$	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_{12} = 4.066 \pm 0.080$	$\hat{\alpha}_{22} = -3.061 \pm 0.060$	$\hat{b}_{12} = -7.119 \pm 0.152$ $\hat{b}_{22} = -4.096 \pm 0.091$ $\hat{b}_{32} = -2.079 \pm 0.068$	$\hat{w}_{12} = 0.300$ $\hat{w}_{22} = 0.300$ $\hat{w}_{32} = 0.400$
k=3	$\hat{\alpha}_{13} = -2.073 \pm 0.041$	$\hat{\alpha}_{23} = 2.062 \pm 0.044$	$\hat{b}_{13} = -5.123 \pm 0.094$ $\hat{b}_{23} = -2.092 \pm 0.038$	$\hat{w}_{13} = 0.599$ $\hat{w}_{23} = 0.401$

Fixed- and random effects coefficients estimated by MSPEM algorithm for the DGP in Eq. (20).

	$\hat{\alpha}_{1k}$	$\hat{\alpha}_{2k}$	$\hat{b}_{m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_{12} = -1.196 \pm 0.036$	$\hat{\alpha}_{22} = -2.775 \pm 0.085$	$\hat{b}_{12} = 4.793 \pm 0.141$ $\hat{b}_{22} = 1.802 \pm 0.069$ $\hat{b}_{32} = -0.103 \pm 0.134$	$\hat{w}_{12} = 0.300$ $\hat{w}_{22} = 0.301$ $\hat{w}_{32} = 0.399$
k=3	$\hat{\alpha}_{13} = -1.673 \pm 0.039$	$\hat{\alpha}_{23} = 1.659 \pm 0.049$	$\hat{b}_{13} = -1.599 \pm 0.056$ $\hat{b}_{23} = -4.788 \pm 0.209$	$\hat{w}_{13} = 0.600$ $\hat{w}_{23} = 0.400$

Fixed- and random effects coefficients estimated by MSPEM algorithm for the DGP in Eq. (21).

	$\hat{\alpha}_k$	$\hat{b}_{1m_k k}$	$\hat{b}_{2m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_2 = -5.008 \pm 0.175$	$\hat{b}_{112} = -5.962 \pm 0.235$ $\hat{b}_{122} = -4.729 \pm 0.128$ $\hat{b}_{132} = -8.023 \pm 0.237$	$\hat{b}_{212} = 5.078 \pm 0.209$ $\hat{b}_{222} = 2.727 \pm 0.121$ $\hat{b}_{232} - 1.183 = \pm 0.087$	$\hat{w}_{12} = 0.300$ $\hat{w}_{22} = 0.300$ $\hat{w}_{32} = 0.400$
k=3	$\hat{\alpha}_3 = 1.996 \pm 0.039$	$\hat{b}_{113} = 0.736 \pm 0.058$ $\hat{b}_{123} = -0.887 \pm 0.054$	$\hat{b}_{213} = -3.642 \pm 0.092$ $\hat{b}_{223} = 2.439 \pm 0.165$	$\hat{w}_{13} = 0.600$ $\hat{w}_{23} = 0.400$

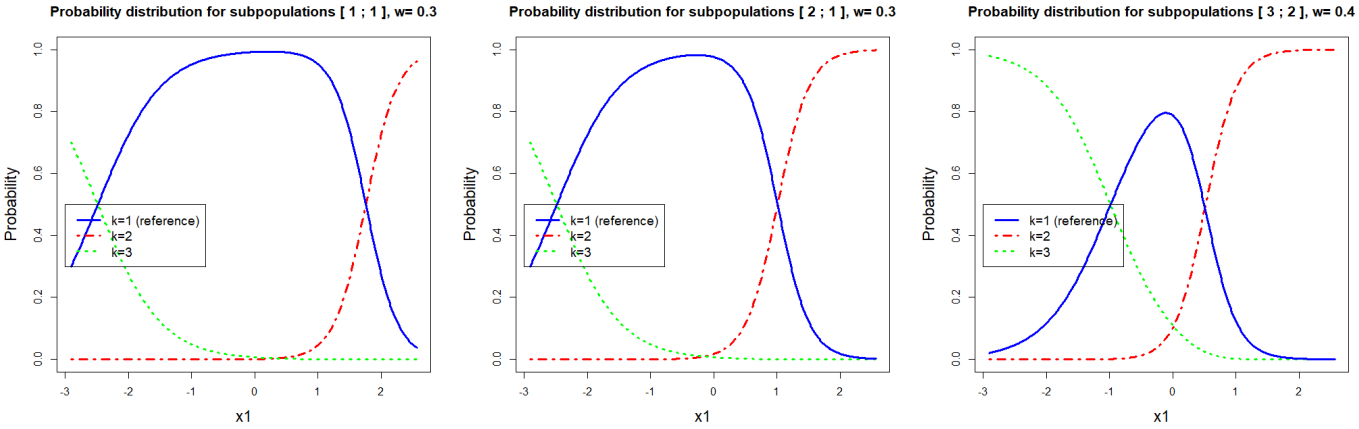
Fixed- and random effects coefficients estimated by MSPEM algorithm for the DGP in Eq. (22).

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 100 runs (randomly chosen). Given the data generating process in Eq. (20), (21) and (22), the joint distribution of the two random effects coefficients distributions has, in all the three settings, three non-zero 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 the average effect of the second covariate[5]. 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

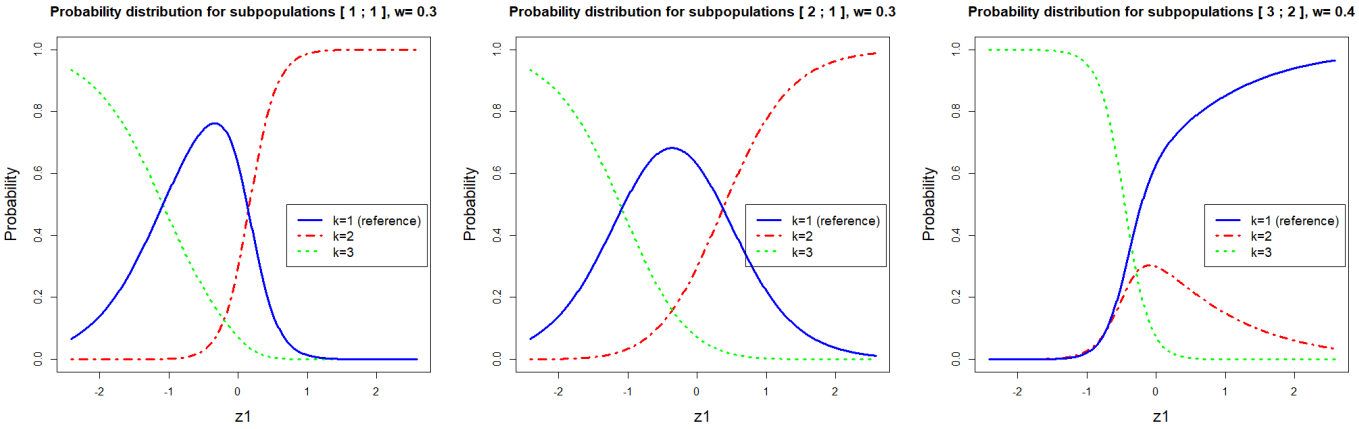
[5]This choice is due to the fact that we are interested in visualizing the trends of the logits for the different values of the random effects coefficients, i.e. the intercept and the slope relative to z_1.

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

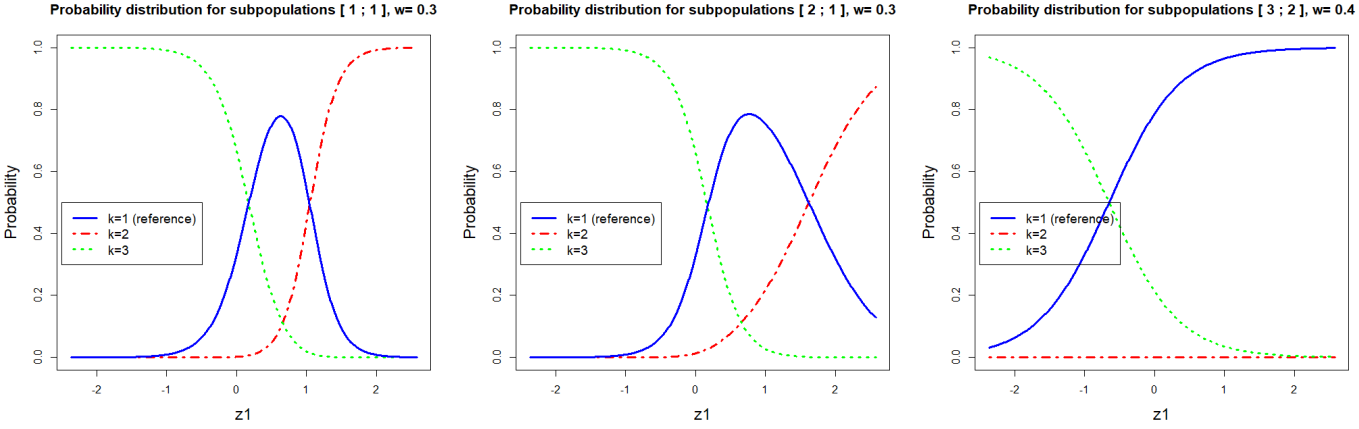
Fig 1: Baseline-category logits estimated by MSPEM algorithm for the three DGPs in Eq. (20), (21) and (22), first, second and third row respectively. Each row reports the three combinations with non-zero weight of the two random effects distributions, i.e. the one relative to category k=2 and the one 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 panel (c), (f) and (i) report the logits for subpopulation 3 relative to category $k = 2$ and subpopulation 2 relative to category $k = 3$.



(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).



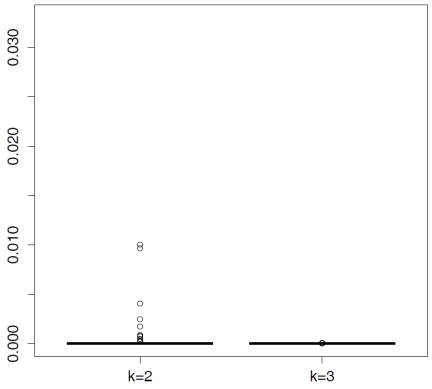
(d) Subpop. $[b_{12}; b_{13}]$ of DGP in (21). (e) Subpop. $[b_{22}; b_{13}]$ of DGP in (21). (f) Subpop. $[b_{32}; b_{23}]$ of DGP in (21).



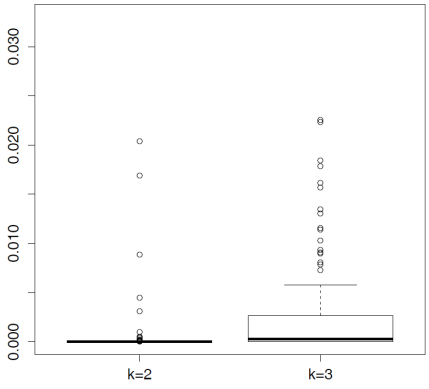
(g) Subpop. $[\mathbf{b}_{12}; \mathbf{b}_{13}]$ of DGP in (22). (h) Subpop. $[\mathbf{b}_{22}; \mathbf{b}_{13}]$ of DGP in (22). (i) Subpop. $[\mathbf{b}_{32}; \mathbf{b}_{23}]$ of DGP in (22).

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, i.e. 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 $E_i = -\sum_{m_k=1}^{M_k} W_{im_kk} \ln(W_{im_kk})$ 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 entropy of the matrices W_k constitutes also a driver for the choice of the tuning parameter D_k, suggesting a lower bound for D_k that minimizes the entropy[6]. Figure 2 reports the distribution of the entropy of W_{i2} and W_{i3}, for $i = 1, \ldots, I$, for the three simulated cases, mediated on the runs in which the algorithm identifies $M_2 = 3$ and $M_3 = 2$.

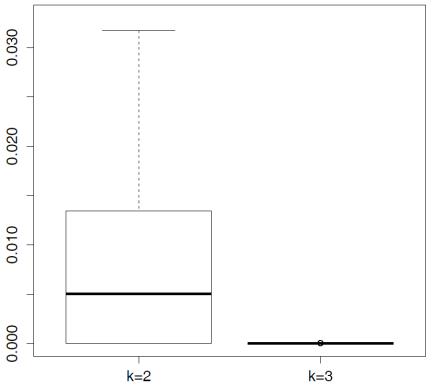
Fig 2: Boxplots of the entropy of W_k, for $k = \{2, 3\}$, measured for each group, obtained by mediating the entropy in the runs in which the algorithm identifies $M_2 = 3$ and $M_3 = 2$, for the random intercept case (a), random slope case (b) and random intercept and slope case (c).



(a) Entropy of matrix W for the random intercept case



(b) Entropy of matrix W for the random slope case



(c) Entropy of matrix W for the random intercept and slope case

We observe that the entropy level is always very low (considering that maximum uncertainty corresponds to the maximum entropy of $-\ln(1/3) = 1.098$ for $k = 2$ and $-\ln(1/2) = 0.693$ for $k = 3$.), 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 increases (and this happens accordingly to the order (a) - only random intercept, (b) - only random slope, and (c) - random intercept

[6]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.

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 paramaters 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.

4. Case study: University student dropout across engineering degree programmes. In the last decades, the analysis of university students dropout is receiving particular attention in the educational context (Aina, 2013; Aljohani, 2016; Belloc et al., 2011). The dropout phenomenon regards those students who conclude their university career without obtaining the degree. Universities are interested in identifying students at risk of dropout, and the determinants of their dropout, in the perspective of understanding the phenomenon and defining new tutoring activities to help them (Aljohani, 2016). Within this context, we present a case study in which we apply the MSPEM algorithm to data about Politecnico di Milano (PoliMi) students, in order to model different categories of students and to identify similar subpopulations of degree courses. This work is within the Student Profile for Enhancing Engineering Tutoring (SPEET) ERASMUS$^+$ project that aims to determine and categorize the different profiles for engineering students across Europe, in order to improve tutoring actions so that they help students to achieve better results and to complete the degree successfully (Barbu et al., 2019). Politecnico di Milano is the largest technical university in Italy and it offers higher education courses in engineering, architecture and design. In our case study, we focus on all concluded careers of students enrolled in an engineering program 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. The aim of this study is to apply the MSPEM algorithm to these data in order to model the dropout probability of students by means of student characteristics and considering their nested structure within degree programmes. In particular, we are interested in analysing whether there are degree programmes in which students are more/less likely to dropout, after adjusting for student characteristics. The MSPEM algorithm permits to identify subpopulations of degree programmes, depending on their effect on students dropout probability.

We exclude form 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 $1,246$ students, respectively). 32.7% of these careers is concluded with a dropout, while the remaining 67.3% regards graduated students. We distinguish among two types of dropout:

- *early dropout* - occurs when the student drops within the 3^{rd} semester after enrolment;
- *late dropout* - occurs when the student drops after the 3^{rd} semester after enrolment.

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 graduated ones. Regarding student characteristics, besides the type of concluded career and the degree program the student is enrolled in, we consider the number of European Credit Transfer System credits (ECTS), i.e. the credits he/she obtained at the first semester of the first year of career (the variable has been standardized in order to have 0 mean and 1 sd) and his/her gender (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 graduated) and guided by the aim of predicting student dropout as soon as possible, i.e. 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., 2020; Pellagatti et al., 2020; Fontana et al., 2018). Table 3 reports the variables considered in the analysis with their explanation.

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

Variable	**Description**	**Type of variable**
`Status`	Type of concluded career	3-levels factor (G = graduated; $D1$ = 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 program the student is enrolled in	19-levels factor

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

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

where

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

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 $2-$dimensional vector of fixed effects covariates, that contains student `Gender` and `TotalCredits1.1`; $\boldsymbol{\alpha}_k$ is the 2-dimensional vector of fixed effects coefficients relative to the $k-$th category; and δ_{ik} is the random intercept relative to the $i-$th degree program (`DegProg`) and to the $k-$th category.

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 sub-population) and $D_k = 0.3$, for $k = 2, 3$. The algorithm converges in 7 iterations and identifies 4 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 see 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

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	$\hat{\alpha}_{13} = -0.685$	$\hat{\alpha}_{23} = -1.899$	$\hat{b}_{13} = -2.152$	$\hat{w}_{13} = 0.210$
			$\hat{b}_{23} = -1.733$	$\hat{w}_{23} = 0.421$
			$\hat{b}_{33} = -1.219$	$\hat{w}_{33} = 0.262$
			$\hat{b}_{43} = -0.880$	$\hat{w}_{43} = 0.107$

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. For each k, subpopulation 1 contains the degree programmes associated to the lowest random intercept, i.e. 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, i.e. degree programmes in which students are more likely to dropout, with respect to the average.

TABLE 5
Distribution of the 19 degree programmes across the 4 identified subpopultions relative to $k = 2, 3$. For each k, the order of the 4 subpopulations is coherent to the one of the estimated random intercepts in Table 4.

Early dropout (k=2)			
Subpopulation 1	Subpopulation 2	Subpopulation 3	Subpopulation 4
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)			
Subpopulation 1	Subpopulation 2	Subpopulation 3	Subpopulation 4
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		

Regarding early dropout (i.e. $k = 2$), from Table 5 we see that the most numerous subpopulation (subpopulation 1, containing 8 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

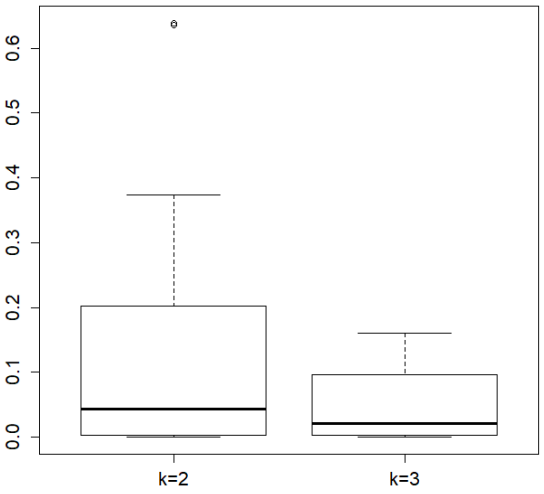
contain degree programmes in which students are more likely to dropout, net to their personal characteristics. In particular, biomedical engineering is identified as an outlier, associated to the highest early dropout probability[7]. For late dropout (i.e. $k = 3$), the most numerous subpopulation is subpopulation 2 (containing 8 degree programmes out of 19 - $\hat{w}_{23} = 0.421$) and, with respect to it, there is a subpopulation of degree programmes associated to a lower late dropout probability and two subpopulations associated to higher late dropout probability. In particular electronic engineering anf engineering of computing system (subpopulation 4) are the ones in which students are more likely to late dropout.

In order to evaluate the uncertainty of classification of the MSPEM algorithm in this case study, panel (a) in Figure 3 reports the entropy distribution of the weight matrices W_k, for $k = 2, 3$. With respect to the maximum entropy of 1.38 relative to the presence of 4 subpopulations, the entropies of W_1 and, especially, W_2 are low (in particular, entropy median and mean are 0.043 and 0.151 for $k = 2$ and 0.021 and 0.048 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[8]. 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 dropout are also less likely to late dropout and vice-versa.

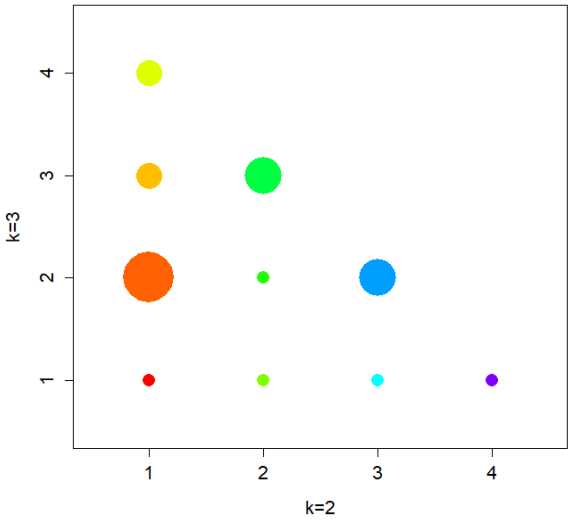
[7] This result in very reasonable since in Italy many students who can not access the medicine faculty, given to the selective entrance exam, attend different faculties, e.g. biomedical engineering, waiting to be admitted to medicine.

[8] Note that for small variations of D around 0.3, e.g. ± 0.05, results do not change.

Fig 3: Panel (a) reports the boxplots of the 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 proportionla to the number of degree programmes belonging to the couple $(m_{\phi 1}, m_{\psi 2})$, for $\phi, \psi = 1, \ldots, 4$.



(a) Entropy of the matrices W_k, for $k = \{2, 3\}$.



(b) Degree programmes distribution across subpopulations.

4.1. *Comparison with MCMCglmm method.* As we said in the Introduction, Hadfield et al. (2010) propose an MCMC method for multi-response generalized linear mixed-models, that provides a robust strategy for marginalizing the random effects. Here, we apply 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 $2,000$. Fixed effects estimates are reported in Table 6, while random intercepts with their confidence intervals are shown in Figure 4. Obviously, since the two methods assume the random effects coefficients to follow different distributions, they lead to different types of results: the MSPEM algorithm identifies a latent structure at the degree programmes level, grouping degree programmes into subpopulations; the MCMCglmm method estimates a single intercept for each degree program, obtaining a ranking of degree courses. Therefore, we are interested in seeing whether the subpopulations identified by the MSPEM algorithm are coherent with the ranking of the MCMCglmm intercepts. 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, results are satisfyingly consistent. For early dropout, Biomedical Engineering, that composes the "outlier" Subpopulation 4, is also the first in the intercepts ranking in panel (a) of Figure 4, resulting to be the degree course associated to the highest early dropout probability by both the methods. Equivalently, Subpopulation 1, associated to the lowest early dropout probability, contains degree courses 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 4 subpopulations identified by the MSPEM algorithm groups the 19 degree courses according to the ranking shown in panel (b) of Figure 4, identifying the first two degree courses, i.e. Engineering of Computing Systems and Electronic Engineering, as components of the subpopulation associated to the highest late dropout probability.

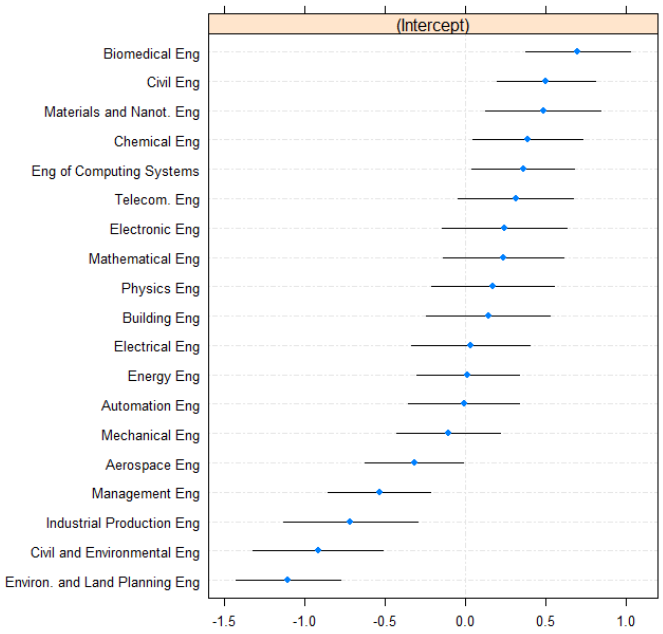
As a final consideration, the identification of subpopulations might be interpreted as a robustness check tool for the groups ranking that we obtain when assuming normal distributed random effects: in the full parametric context, we have no evidence to document differences and, consequently, to create 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.

TABLE 6
Fixed effects estimates of the MCMCglmm *method.*

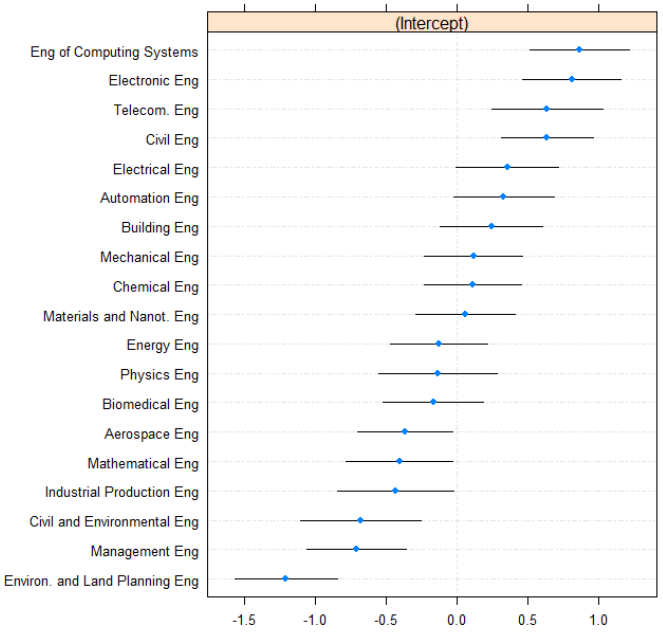
	Variable name	post.mean	$l-95\%$ CI	$u-95\%$ CI	pMCMC
k=2	`Intercept`	-2.552	-2.854	-2.269	$<0.001**$
	`Gender`	-0.027	-0.106	0.153	0.769
	`TotalCredits1.1`	-2.797	-2.884	-2.702	$<0.001**$
k=3	`Intercept`	-2.354	-2.672	-2.049	$<0.001**$
	`Gender`	-0.634	-0.464	-0.802	$<0.001**$
	`TotalCredits1.1`	-2.135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

Note: *p<0.1; **p<0.05; ***p<0.01

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.



(a) Degree programmes intercepts for Early Dropout



(b) Degree programmes intercepts for Late Dropout

5. Concluding remarks and future perspectives. This paper proposes a semiparametric multinomial mixed-effects linear model, together with an Expectation-Maximization algorithm to estimate its parameters. 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 the identification of a latent structure at the higher level of the hierarchy in which groups collapse into a finite and *a priori* unknown number of subpopulations. In particular, we identify 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 to express the marginal distribution of the response as a weighted sum, avoiding difficult integration problems.

We show a simulation study in which we test the algorithm for different random effects configurations, proposing a way to evaluate the method performance.

Lastly, we apply the proposed algorithm to a real case study in which we model different profiles of engineering university students, considering 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 then compare our results with the ones obtained by applying a full parametric method, the *MCMCglmm*, to the same case study, underlining similarities and differences and exploiting the different types of results provided by parametric and semiparametric methods.

This work enters in the literature about mixed-effects models with discrete random effects (Aitkin, 1999; Hartzel, 2000; Masci et al., 2019), proposing a novel method that deals with multinomial responses. Several issues are still unresolved and further developments are needed regarding the random effects structure assumptions. At the current state of the art, random effects are assumed to be independent across categories k, for $k = 2, \dots, K$. Since the random effects from different logits arise from the same subjects, this assumption may be unrealistic. Our first future perspective is therefore to extend the proposed method to deal with more complex dependence structures of the random effects across categories.

APPENDIX A: PROOF OF THE INCREASING LIKELIHOOD PROPERTY OF THE MSPEM ALGORITHM

In appendix A of Azzimonti et al. (2013), the authors prove the increasing likelihood property of the EM algorithm which we are inspired by. In their paper, they propose a nonparametric mixed-effects model for unsupervised classification for a continuous response that might be non-linear, but with density function in the exponential family. Their response variable, considering our notation, is modelled as:

$$\begin{aligned} \mathbf{y}_i &= f(\boldsymbol{\alpha}, \boldsymbol{\delta}_i) + \boldsymbol{\epsilon}_i \qquad && i = 1, \dots, I \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{1}_n) && i.i.d. \end{aligned}$$

and they also assume that the random effects coefficients follow a discrete distribution with M mass points $(\mathbf{b}_1, \dots, \mathbf{b}_M)$ with associated weights $(w_1, \dots, w_M)$. They prove that the parameters estimates obtained by maximizing their likelihood, thanks to its convexity property, can be computed in two separate steps: one for computing the weights of the discrete distribution of the random effects and one for computing fixed effects coefficients and random effects support points iteratively. In particular, they prove that the updated parameters are obtained such that:

$$L(\boldsymbol{\alpha}^{(up)}, \sigma^2(up)|\mathbf{y}) \geq L(\boldsymbol{\alpha}, \sigma^2|\mathbf{y}),$$

where $\boldsymbol{\alpha}^{(up)}$ and $\sigma^2(up)$ are the updated fixed effects coefficients and error variance and the likelihood $L(\boldsymbol{\alpha}^{(up)}, \sigma_2^{(up)}|\mathbf{y})$ is computed summing up the random effects with respect to the updated discrete distribution $(\mathbf{b}_m^{(up)}, w_m^{(up)})$ for $m = 1, \dots, M$. Following the steps presented in appendix A of Azzimonti et al. (2013), we observe that, thanks to the definition of the likelihood function in Eq. (9), we have that:

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

All these terms can be bounded above by the quantity:

$$\log\left(\frac{p(\mathbf{y}_i|\boldsymbol{A}^{(up)})}{p(\mathbf{y}_i|\boldsymbol{A})}\right) \geq Q_i(\theta^{(up)}, \theta) - Q_i(\theta, \theta), \tag{26}$$

where

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

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

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

we obtain, thanks to Eq. (26), an upper bound for the quantity of interest

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

In order to show now that $\forall\theta, Q(\theta^{(up)},\theta) \geq Q(\theta,\theta)$, we can show that, $\forall\theta$ fixed, $\theta^{(up)}$ is defined as the $\arg\max_{\tilde{\theta}} Q(\tilde{\theta},\theta)$.
Defining W_{im} as the probability that the $i-$th group belongs to the $m-$th combination among the M_{tot} possible combinations, conditionally on the observations $\mathbf{y}_i$ and given the fixed effects parameters $\boldsymbol{A}$, we obtain

$$\begin{aligned}
Q(\tilde{\theta},\theta) &= \sum_{i=1}^{I}\sum_{m=1}^{M}\left(\frac{w_m p(\mathbf{y}_i|\boldsymbol{A},\mathbf{B}_m)}{p(\mathbf{y}_i|\boldsymbol{A})}\right)\log(\tilde{w}_m p(\mathbf{y}_i|\tilde{\boldsymbol{A}},\tilde{\mathbf{B}}_m)) = \\
&= \sum_{i=1}^{I}\sum_{m=1}^{M} W_{im}\log(\tilde{w}_m p(\mathbf{y}_i|\tilde{\boldsymbol{A}},\tilde{\mathbf{B}}_m)) = \\
&= \sum_{i=1}^{I}\sum_{m=1}^{M} W_{im}\log(\tilde{w}_m) + \sum_{i=1}^{I}\sum_{m=1}^{M} W_{im}\log(p(\mathbf{y}_i|\tilde{\boldsymbol{A}},\tilde{\mathbf{B}}_m)) = \\
&= J_1(\tilde{w}_1,\ldots,\tilde{w}_{M_{tot}}) + J_2(\tilde{\boldsymbol{A}},\tilde{\mathbf{B}}_1,\ldots,\tilde{\mathbf{B}}_M). \qquad (27)
\end{aligned}$$

The functionals J_1 and J_2 can be maximized separately. In particular, by maximizing the functional J_1 we obtain the updates in (11) for the weights of the random effects distribution and, by maximizing the functional J_2 in an iterative way, we obtain the estimates of $\boldsymbol{A}$ and $\mathbf{B}_m$, for $m = 1,\ldots,M_{tot}$, in (13). In particular, assuming that the random effects distributions relative to each category k, for $k = 2,\ldots,K$, are independent, the maximization of the functionals J_1 and J_2 can be done separately for each category k. Indeed, under the independence assumption:

$$\tilde{w}_m = \prod_{k=2}^{K} \tilde{w}_{mk} \qquad m = 1,\ldots,M_{tot},$$

where $\tilde{w}_{mk}$ is the $k-$relative weight that contributes to form the $m-$th combination. Finding the maximum of $\tilde{w}_m$ is equivalent to find the maximum of the $(K-1)$ components whose product compose it. Equivalently, thanks to the convexity of the logarithm, the maximization of J_2 can be done separately for each category k.

APPENDIX B: TECHNICAL DETAILS ABOUT THE MSPEM ALGORITHM

In this section, we give some insights about the discrete distribution support points initialization, the support points collapse criteria and the convergence criteria.

The EM algorithm starts considering an equal number of mass points[9] relative to each category, $M_k^* = M^* = I$, for $k = 2, \dots K$. Given this number of support points, the initialization of the support points is done in the following way:

a) Random effects: the starting I support points are obtained by fitting a simple multinomial logistic regression within each group, i.e. without considering the nested structure but considering I distinct models, and estimating the parameters - relative to the covariates that will be considered part of the random effects in the mixed-effects model - for each of the I groups, relative to each category k, for $k = 2, \dots, K$. The weights are uniformly distributed on these I support points.
b) Fixed effects: the starting values of $\boldsymbol{A}$ are computed as the mean of the coefficients - relative to the covariates that will be considered part of the fixed effects in the mixed-effects model - obtained by the I multinomial logistic models. In particular, for each category k, for $k = 2, \dots, K$, we obtain $\boldsymbol{\alpha}_k = 1/I \sum_{i=1}^{I} \boldsymbol{\alpha}_{ik}$.

Nonetheless, if the number of groups I is extremely large, the elevated number of support points makes the algorithm relatively slow and this is not strictly necessary. In this case, following the method proposed in Masci et al. (2019), we rescale the initialization of the support points of the $(K-1)$ random effect distributions in the following way:

- we choose a reasonable number $M_k^* = M^* < I$;
- we extract M^* points, for each category k, $k = 2, \dots, K$, from a uniform distribution with support on the entire range of possible values, i.e. the range of the distribution of coefficients obtained by fitting I distinct multinomial logistic regressions;
- we uniformly distribute the weights on these M^* support points, for each $k = 2, \dots, K$.

During the iterations, the EM algorithm performs the reduction of the support points of the random effects discrete distributions, in order to identify, for $k = 2, \dots, K$, $M_k* < I$ subpopulations that describe the latent structure relative to each contrast $k' = k - 1$. To this end, we fix a threshold distance D_k, for $k = 2, \dots, K$, and when two mass points, relative to category k, $\mathbf{b}_{m_k k}$ and $\mathbf{b}_{m_l k}$ are closer than D_k, they collapse to a unique point $\mathbf{b}_{m_{k,l} k} = (\mathbf{b}_{m_k k} + \mathbf{b}_{m_l k})/2$ with associated weight $w_{m_{l,k} k} = w_{m_l k} + w_{m_k k}$. The threshold D_k is allowed to differ across the categories, i.e. 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, $k = 2, \dots, K$, 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 $k = 2, \dots, K$ and we remove mass points with weight $w_{m_k k} < \tilde{w}_k$. 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 $k = 2, \dots, K$, will result as anomalous groups.

D_k and $\tilde{w}_k$ are two tuning parameters that tune the estimates of the subpopulations. The choice of D_k depends on how much we want to be sensitive to the differences between subpopulations: the higher is D_k, the lower is the number of subpopulations and the less homogeneous are the groups within subpopulations. D_k depends also on the order of magnitude of the data and its choice is driven by the range of the distribution of coefficients obtained by fitting I distinct multinomial logistic regressions (described in the initialization of the support points). The choice of $\tilde{w}_k$ depends on the minimum number of groups that we allow within

[9] Alternatively, it is possible to choose different starting numbers of mass points for each category k, for $k = 2, \dots K$. This choice is arbitrary.

each subpopulation. When one or more mass points are deleted, the remaining weights are reparameterized in such a way that they sum up to 1.

At each iteration of the EM algorithm, given the estimated number of mass points, we estimate all the parameters in Eq. (7) in an iterative way, updating the coefficients of both fixed and random effects of each contrast, until convergence or until we reach the maximum number of subiterations fixed a priori for this stage, `itmax`. At the beginning of the iterative process, the algorithm performs the dimensional reduction of the mass points on the basis of only the distance between the mass points. When the estimates are stable, meaning that all the differences between the estimates of the parameters at two consecutive iterations are smaller than fixed tolerance values, or after a given number of iterations, `it1`, the algorithm continues performing the dimensional reduction of the support points on the basis of also the criterion of the minimum weight $\tilde{w}_k$. Convergence is finally reached when all the differences between the estimates of the parameters in two consecutive iterations are smaller than fixed tolerance values. In particular, we fix the tolerance values for the estimates of both the parameters of fixed and random effects to `tollF` and `tollR` respectively, which depend on the scale of the parameters[10].

[10]More details about the choice of the tolerance values and the tuning parameters can be found in Masci et al. (2019).

REFERENCES

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

Aina, C. (2013). Parental background and university dropout in italy. *Higher Education*, 65(4):437–456.

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

Aljohani, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher education studies*, 6(2):1–18.

Anderson, C. J., Kim, J.-S., and Keller, B. (2013). Multilevel modeling of categorical response variables. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, pages 481–519.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(2):203–210.

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

Barbu, M., Vilanova, R., Vicario, J., Pereira, M. J., Alves, P., Podpora, M., Kawala-Janik, A., Prada, M., Dominguez, M., Spagnolini, A., 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. *Journal of applied Statistics*, 38(10):2225–2239.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

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

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., and Paganoni, A. (2020). Not the magic algorithm: modelling and early-predicting students dropout through machine learning and multilevel approach.

Coull, B. A. and Agresti, A. (2000). Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics*, 56(1):73–80.

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

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

Dos Santos, D. M. and Berridge, D. M. (2000). A continuation ratio random effects model for repeated ordinal responses. *Statistics in medicine*, 19(24):3377–3388.

Fontana, L., Masci, C., Ieva, F., and Paganoni, A. (2018). Performing learning analytics via generalized mixed-effects trees.

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

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

Hadfield, J. D. et al. (2010). Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package. *Journal of Statistical Software*, 33(2):1–22.

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

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

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

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *The annals of statistics*, pages 86–94.

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

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

McCulloch, C., Lin, H., Slate, E., and Turnbull, B. (2002). Discovering subpopulation structure with latent class mixed models. *Statistics in medicine*, 21(3):417–429.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89(425):330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170.

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

Muthén, B. (2004). Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences*, 345(368):106–109.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods*, 4(2):139.

Pellagatti, M., Masci, C., Ieva, F., and Paganoni, A. (2020). Generalized mixed-effects random forest: a flexible approach to predict university student dropout.

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

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

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. *Journal of computational and Graphical Statistics*, 9(1):141–157.

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

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

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.

Wolfinger, R. and O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, 48(3-4):233–243.

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006). General design bayesian generalized linear mixed models. *Statistical science*, pages 35–51.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

42/2020 Miglio, E.; Parolini, N.; Quarteroni, A.; Verani, M.; Zonca, S.
A spatio-temporal model with multi-city mobility for COVID-19 epidemic

43/2020 Quarteroni, A.; Vergara, C.
Modeling the effect of COVID-19 disease on the cardiac function: A computational study

39/2020 Martinolli, M.; Biasetti, J.; Zonca, S.; Polverelli, L.; Vergara, C.
Extended Finite Element Method for Fluid-Structure Interaction in Wave Membrane Blood Pumps

40/2020 Fresca, S.; Manzoni, A.; Dedè, L.; Quarteroni, A.
Deep learning-based reduced order models in cardiac electrophysiology

41/2020 Cannistrà,M.; Masci, C.; Ieva, F.; Agasisti, T.; Paganoni, A.M.
Not the magic algorithm: modelling and early-predicting students dropout through machine learning and multilevel approach

38/2020 Sollini, M.; Kirienko, M.; Cavinato, L.; Ricci, F.; Biroli, M.; Ieva, F.; Calderoni, L.; Tabacchi,
Methodological framework for radiomics applications in Hodgkin's lymphoma

34/2020 Antonietti, P.F.; Mazzieri, I.; Nati Poltri, S.
A high-order discontinuous Galerkin method for the poro-elasto-acoustic problem on polygonal and polyhedral grids

35/2020 Morbiducci, U.; Mazzi, V.; Domanin, M.; De Nisco, G.; Vergara, C.; Steinman, D.A.; Gallo, D
Wall shear stress topological skeleton independently predicts long-term restenosis after carotid bifurcation endarterectomy

36/2020 Pellagatti, M.; Masci, C.; Ieva, F.; Paganoni A.M.
Generalized Mixed-Effects Random Forest: a flexible approach to predict university student dropout

37/2020 Fumagalli, A.; Scotti, A.
A mathematical model for thermal single-phase flow and reactive transport in fractured porous media