Submitted to the Annals of Applied Statistics

A MULTINOMIAL MIXED-EFFECTS MODEL WITH DISCRETE RANDOM EFFECTS FOR MODELLING DEPENDENCE ACROSS RESPONSE CATEGORIES

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*

We propose a Semi-Parametric Mixed-Effects Multinomial regression model to deal with estimation and inference issues in the case of categorical and hierarchical data. The proposed modelling assumes the probability of each response category to be identified by a set of fixed and random effects parameters, estimated by means of an Expectation-Maximization algorithm. Random effects are assumed to follow a discrete distribution with an *a priori* unknown number of support points. For a $K-$category response, this method identifies a latent structure at the highest level of grouping, where groups are clustered into $(K-1)-$dimensional subpopulations. This method is an extension of the multinomial semi-parametric EM algorithm proposed in the literature, in which we relax the independence assumption across random-effects relative to different response categories. Since the category-specific random effects arise from the same subjects, their independence assumption is seldom verified in real data. In this sense, the proposed method properly models the natural data structure, as emerges by the results of simulation and case studies. The obtained results prove the importance of taking into account the data dependence structure in real data applications.

1. Introduction & Literature. The big data hera has increased the collection of any type of data and, among the others, of categorical data. Quality of life, patient recovery or pain, diagnostic evaluations, political or religious philosophy, educational evaluations are only few examples of ordered and unordered categorical data that are continuously collected and used by companies, countries or different stakeholders to enhance their work. In the framework of generalized linear models, multinomial responses have always been treated separately from other response distributions. Despite the multinomial distribution belongs to the exponential family, most of the algorithms and procedures that implement linear models for responses in the exponential family do not include the multinomial one. Since, for a multicategory response, multiple logits must be considered, multinomial models can be better treated as multivariate generalised linear models (Tutz and Hennevogl, 1996). This applies also in the context of hierarchical data, i.e. data containing observations naturally nested within groups, such as longitudinal data or repeated measurements (Agresti, 2018). Hierarchical data are usually treated by means of mixed-effects models (Pinheiro and Bates, 2006), but their developments to handle multinomial responses are also quite limited.

Mixed-effects models include in the linear predictor both *fixed effects* associated to the entire population and *random effects* associated to the groups in which observations are nested, drawn at random from the population (Goldstein, 2011). This mechanism allows to account for various 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 are intended for grouped data in which the response variable is continuous (Pinheiro and Bates, 2006). When

Keywords and phrases: Discrete random effects, Multinomial regression, Unsupervised clustering, Multivariate statistics, Higher education.

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 and the marginal distribution of the response can be obtained by integrating out the random effects. 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 little development for multinomial responses. 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 required by their modelling. Indeed, an appropriate link function for nominal responses is the baseline-category logit, where fixed and random coefficients vary according to the response category. For this reason, mixed-effects linear models for a multinomial response are 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). 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). 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 on this topic (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 prior distributions on the parameters. 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.

A more recent branch of the literature about mixed-effects linear models proposes a semi-parametric approach in which the random effects are assumed to follow a discrete distribution with an *a priori* unknown number of support points (Aitkin, 1999; Masci et al., 2019b). While parametric mixed-effects models usually identify a normal distribution of random effects and each highest level unit's point estimate is extracted from this distribution, the semi-parametric approach identifies a classification of highest level units that are clustered into subpopulatons standing on the similarities of their effects. Semi-parametric Mixed-effects Linear Models (SLMMs) have been proposed for a continuous response (Masci et al., 2019b), for multiple continuous responses (Masci et al., 2019a) and for a binary response (Maggioni, 2020). A very recent work proposes semi-parametric mixed-effects linear models for a multinomial response (Masci et al., 2021). The authors in Masci et al. (2021) face a classification problem with hierarchical data. They aim to profile engineering students of Politecnico di Milano (PoliMI) into three categories (early dropout, late dropout and graduated), given some student personal and career information and considering their nested structure within 19 engineering degree programmes. To this end, they propose the Multinomial Semi-Parametric Expectation-Maximization algorithm, called *MSPEM*, for SLMMs dealing with a multinomial response. In particular, they assume the random effects of the model to follow a multivariate discrete

distribution with an *a priori* unknown number of support points, that is allowed to differ across response categories. The discrete distribution assumption on the random effects in a multinomial model allows to express the likelihood as a weighted sum instead of a multiple integral, significantly simplifying the estimation procedure of the model parameters. The baseline-category logit approach considers category-specific fixed and random effects parameters. In Masci et al. (2021), the authors develop a method that assumes the random effects relative to different response categories to be independent. This assumption simplifies the parameters estimation procedure, but it is a strong and often unrealistic assumption, since the random effects of different logits arising from the same subject are expected to be somehow related.

In this work, we propose a new method to fit semi-parametric multinomial mixed-effects models that does not assume the independence across the category-specific random effects distributions and that is able to model different dependence structures across the multinomial categories. In particular, for a multinomial response assuming K different categories, we assume the random effects to follow a joint $(K-1)$-variate discrete distribution. Each of the $(K-1)$ marginal distributions is allowed to have a different number of support points. For this reason, we refer to this proposed method as the Joint Multinomial Semi-Parametric EM algorithm, the *JMSPEM* algorithm. This modelling allows to take into account the dependence structure among the categories when estimating the parameters, resulting in two main advantages: the former is that we avoid bias in the estimates, induced by the natural dependence across categories; the latter is that, by jointly estimating the highest level units effects on the $K-1$ logits, we better investigate and interpret their trends. The assumption of discrete random effects provides a new interpretation of units at the highest level of the hierarchy, that are clustered into subpopulations, identified by the support points of the discrete distribution. This approach has several advantages (Rights and Sterba, 2016): first of all, by assuming discrete random effects, it is possible to identify a latent structure at the highest level of the hierarchy, that is a valuable alternative to the ranking provided by assuming gaussian random effects; secondly, since the semi-parametric approach is more flexible and it does not assume *a priori* any parametric distribution, it can potentially estimate the real distribution of the random effects; third, when the number of groups is extremely large, the identification of subpopulations might help in interpreting the results, thanks to the dimensional reduction; lastly, the identification of subpopulations gives insights about the outlier identification, since the most populated subpopulations reveal which are the reference trends, while the smallest ones contain those groups whose observations tend to have anomalous behaviours with respect to the majority.

In order to insert the *JMSPEM* method in a clear inferential framework, we complete the algorithm by adding a method to compute the standard errors of the estimates and to assess the significance of the coefficients. In particular, regarding the standard errors, the variance of the ML estimator is calculated by the inverse of the Information matrix. For what concerns the significance, for fixed-effects, we assess it by using the likelihood ratio test (Dempster et al., 1977), while, for random-effects, the Variance Partition Coefficient (VPC) for a semi-parametric multinomial mixed-effects model is proposed.

Modelling the dependence across categories increases the dimensional complexity of the estimates, requiring a not trivial computational improvement. In order to test and investigate the potential of the *JMSPEM* algorithm, we retrace both the simulation and the case studies proposed in Masci et al. (2021), underlying the differences and, potentially, the advantages of the *JMSPEM* algorithm with respect to the *MSPEM* one. Results of the simulation study show that the parameters estimated by the *JMSPEM* algorithm have reduced bias and variance with respect to the ones estimated by *MSPEM* algorithm. Moreover, in the case study, the subpopulations identified at the highest level of the hierarchy are robustly coherent with the ranking estimated by the parametric *MCMCglmm* method.

The remaining part of the paper is organized as follows: in Section 2 we describe the *JMSPEM* model and algorithm; in Section 3 we retrace the simulation study proposed in Masci et al. (2021) comparing the results of *JMSPEM* and *MSPEM*; in Section 4 we apply the *JMSPEM* algorithm to the Politecnico di Milano case study presented in Masci et al. (2021) for modelling higher education student dropout and we compare the results obtained by applying three methods: *MSPEM*, *JMSPEM* and the parametric method *MCMCglmm*; Section 5 draws the conclusions.

2. Methodology: joint semi-parametric mixed-effects model for a multinomial response. In this section, we first recall the basics of a mixed-effects multinomial model with discrete random effects (Subsection 2.1), and, then, we present the *JMSPEM* model and algorithm (Subsection 2.2).

2.1. *Mixed-effects models for a multinomial response with discrete random effects.* 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 $j = 1, \ldots, n_i$, is nested within a group i, for $i = 1, \ldots, I$. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ be the n_i-dimensional response vector for observations within the $i-$th group. The multinomial distribution with K categories relative to Y_{ij} is the following:

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

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

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

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

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

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

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

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

The Maximum Likelihood Estimation (MLE) method allows to estimate the model parameters of this probability distribution.

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

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

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

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

In the semi-parametric approach presented in Masci et al. (2021), the coefficients of the random effects are assumed to follow a discrete distribution with an *a priori* unknown number of support points (Masci et al., 2019b). Under this assumption, the multinomial logit takes the form:

$$\eta_{ijk} = \mathbf{x}'_{ij}\boldsymbol{\alpha}_k + \mathbf{z}'_{ij}\mathbf{b}_{m_k k} \qquad m_k = 1, \ldots, M_k, \quad k = 2, \ldots, K, \tag{6}$$

where M_k is the total number of support points of the discrete distribution of $\mathbf{b}$ relative to the $k-$th category, for $k = 2, \ldots, K$. 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 $w_{m_k} \geq 0$:

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

The discrete distributions P_k^*, for $k = 2, \ldots, K$, belong to the class of all probability measures on $\mathcal{R}^q$ and are assumed to be independent. P_k^* is a discrete measure with M_k support points that can then be interpreted as the mixing distribution that generates the density of the stochastic model in (6). 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 semi-parametric maximum likelihood estimator of a mixing distribution, in the case of exponential family densities.

Given this formulation, in Masci et al. (2021) the authors propose the *MSPEM* 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. In the *MSPEM* steps, under the independence assumption across the contrast-specific random-effects, when estimating the support points relative to each contrast, the other contrast-specific random-effects parameters are fixed to the mean of the relative discrete distributions. In other words, when estimating the random effects of a group with respect to a response category, the random effects of this specific group with respect to the other categories are ignored. This assumption simplifies the parameters estimation procedure, but, as previously discussed, it is often too strict and unrealistic.

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

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

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

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

Under these assumptions, the *JMSPEM* parameters estimates can be obtained by maximizing the likelihood in Eq. (8). Thanks to the likelihood convexity property, the maximization can be computed in two separate steps: one for computing the weights of the multivariate discrete distribution of the random effects and one for computing fixed effects coefficients and random effects support points iteratively.

In particular, the updated parameters are obtained such that:

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

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

$$\log\left(\frac{p(\mathbf{y}_i|\boldsymbol{A}^{(up)})}{p(\mathbf{y}_i|\boldsymbol{A})}\right) \geq Q_i(\theta^{(up)}, \theta) - Q_i(\theta, \theta), \tag{9}$$

where

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

$Q_i(\theta, \theta)$ is analogously defined and $\theta = (\boldsymbol{A}, \mathbf{B}_1, \dots, \mathbf{B}_{M_{tot}}, w_1, \dots, w_M)$. 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. (9), 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

$$\begin{aligned}
Q(\tilde{\theta}, \theta) =& \sum_{i=1}^{I} \sum_{m=1}^{M_{tot}} \left(\frac{w_m p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{B}_m)}{p(\mathbf{y}_i|\boldsymbol{A})}\right) \log(\tilde{w}_m p(\mathbf{y}_i|\tilde{\boldsymbol{A}}, \tilde{\mathbf{B}}_m)) = \\
=& \sum_{i=1}^{I} \sum_{m=1}^{M_{tot}} W_{im} \log(\tilde{w}_m p(\mathbf{y}_i|\tilde{\boldsymbol{A}}, \tilde{\mathbf{B}}_m)) = \\
=& \sum_{i=1}^{I} \sum_{m=1}^{M_{tot}} W_{im} \log(\tilde{w}_m) + \sum_{i=1}^{I} \sum_{m=1}^{M} W_{im} \log(p(\mathbf{y}_i|\tilde{\boldsymbol{A}}, \tilde{\mathbf{B}}_m)) = \\
=& J_1(\tilde{w}_1, \dots, \tilde{w}_{M_{tot}}) + J_2(\tilde{\boldsymbol{A}}, \tilde{\mathbf{B}}_1, \dots, \tilde{\mathbf{B}}_{M_{tot}}).
\end{aligned} \tag{10}$$

$\mathbf{w}$ is an array with $K-1$ dimensions, i.e. a $(M_2 \times M_3 \times \ldots \times M_K)-$ dimensional array, and each element w_m represents the weight of the $m-$th $(K-1)-$variate support point. Equivalently, W is an array with K dimensions, i.e. a $(I \times M_2 \times M_3 \times \ldots \times M_K)-$ dimensional array of conditional weights[3]. In particular,

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

and

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

By marginalizing W with respect to k, we obtain the marginal conditional weights matrices W_k, for $k = 2, \ldots, K$. The functionals J_1 and J_2 can be maximized separately. In particular, by maximizing the functional J_1 we obtain the updates for the weights of the random effects distribution and, by maximizing the functional J_2 in an iterative way, we obtain the estimates of $\boldsymbol{A}$ and $\mathbf{B}_m$, for $m = 1, \ldots, M_{tot}$.

The EM algorithm for the maximization of the two functionals 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. The EM algorithm stops when the convergence or a maximum number of iterations are reached. In particular, the update for the parameters is given by:

$$w_m^{(up)} = \frac{1}{I} \sum_{i=1}^{I} W_{im} \qquad m = 1, \ldots, M_{tot}, \tag{13}$$

and

$$(\boldsymbol{A}^{(up)}, \mathbf{b}_1^{(up)}, \ldots, \mathbf{b}_{(K-1)}^{(up)}) = \arg\max_{\boldsymbol{A}, \mathbf{B}_m} \sum_{m=1}^{M_{tot}} \sum_{i=1}^{I} W_{im} \times \ln p(\mathbf{y}_i|\boldsymbol{A}, \mathbf{B}_m). \tag{14}$$

The weight $w_m^{(up)}$ is the mean over the I groups of their conditional weights relative to the $m-$th $(K-1)-$variate support point. Coefficient W_{im} represents the probability that group i belongs to the $m-$th $(K-1)-$variate subpopulation, identified by the relative $K-1$

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

marginal subpopulations, conditionally on observations $\mathbf{y}_i$ and given the fixed coefficients $\boldsymbol{A}$. The maximization step in Eq. (14) involves two steps and it is done iteratively. In the first step, thanks to the convexity of the logarithm, for each category k, for $k = 2, \ldots, K$, we compute the $\arg\max$ with respect to the support points $\mathbf{b}_{m_k k}$, for $m_k = 1, \ldots, M_k$, keeping $\boldsymbol{A}$ and $\mathbf{b}_l$, for $l \neq k$, fixed to the values computed at the previous iteration. In this way, we can maximize the expected log-likelihood (computed in the expectation step) with respect to all support points $\mathbf{b}_{m_k k}$ separately, i.e.

$$\mathbf{b}_{m_k k}^{(up)} = \arg\max_{\mathbf{b}_k} \sum_{c_{m_k}=1}^{C_{tot,k}} \sum_{i=1}^{I} W_{i c_{m_k}} \ln p(\mathbf{y}_i | \boldsymbol{A}, \mathbf{b}_k, \mathbf{B}_{c_{m_k}}^{(old)})$$

$$m_k = 1, \ldots, M_k, \quad k = 2, \ldots, K. \tag{15}$$

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

$$p(\mathbf{y}_i | \boldsymbol{A}, \mathbf{b}_k, \mathbf{B}_{c_{m_k}}^{(old)}) = \prod_{j=1}^{n_i} \prod_{\gamma=1}^{K} \left(\frac{\exp(\eta_{ij\gamma})}{1 + \sum_{\nu=2}^{K} \exp(\eta_{ij\nu})} \right)^{\{\mathbf{1}_{y_{ij}=\gamma}\}}, \tag{16}$$

where

$$\eta_{ij\gamma} = \begin{cases} \mathbf{x}_{ij}' \boldsymbol{\alpha}_k + \mathbf{z}_{ij}' \mathbf{b}_k & \text{if} \quad \gamma = k \\ \mathbf{x}_{ij}' \boldsymbol{\alpha}_\gamma + \mathbf{z}_{ij}' \mathbf{b}_{(m_\gamma \gamma)_{c_{m_k}}}^{(old)} & \text{if} \quad \gamma \neq k \end{cases}. \tag{17}$$

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

In the second step, we fix the support points of the random effects distributions computed in the previous step and we compute the $\arg\max$ in Eq. (14) with respect to $\mathbf{A}$. Again, thanks to the convexity of the logarithm, we can compute the $\arg\max$ in Eq. (14) with respect to $\boldsymbol{\alpha}_k$, separately for each $k = 2, \ldots, K$, keeping $\boldsymbol{\alpha}_l$, for $l \neq k$ fixed to the values computed at the previous iteration.

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

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

$$i = 1, \ldots, N. \tag{18}$$

Notice that, despite *MSPEM* and *JMSPEM* algorithm skeletons are basically the same, substantial differences regard essentially the estimation of the random effects, i.e. of the

weights (Eq. 13) and of the random effects support points (Eq. 15). In the *MSPEM* algorithm, only marginal weights and marginal conditional weights matrices are treated and in the maximization in Eq. 15, the groups' belonging to the subpopulations relative to the other categories are ignored. In the *JMSPEM* algorithm, all weights and conditional weights are treated in their multivariate setting and the function to be maximized in Eq. 15 takes into account the conditional weights of groups across all categories. The multivariate optimization implies an increased computational cost, that scales with the number of covariates and of response-categories.

During the iterations of the EM algorithm, the reduction of the support points of the random effects discrete distributions is performed. All details about the discrete distribution support points initialization, the support points collapse criterion, the convergence criterion and model identifiability can be found in Masci et al. (2021).

Besides the point estimates of both fixed and random effects coefficients, a further improvement provided by the JMSPEM algorithm regards the computation of their standard errors and the assessment of their significance. The variance of maximum likelihood estimators is calculated by the inverse of the Information matrix. Considering θ the parameter to be estimated by a ML method:

$$\begin{aligned} var(\theta) &= [I(\theta)^{-1}] \\ &= (-E[H(\theta)])^{-1} \\ &= \left(-E\left[\frac{\partial^2 ln\mathcal{L}(\theta}{\partial\theta\partial\theta'}\right]\right)^{-1} \end{aligned}$$

where $H(\theta)$ is the Hessian matrix, i.e., the matrix of second derivatives of the likelihood $\mathcal{L}$ with respect to the parameter θ, and $E[H(\theta)]$ is its expected value. The standard error of each estimator is just the square root of this estimated variance (King, 1989; Long and Long, 1997). Moreover, the fixed-effects coefficients significance is assessed by means of the likelihood ratio test (Agresti, 2018).

3. Simulation study. In this section, we retrace the simulation study proposed in Masci et al. (2021), to compare the performances of the *JMSPEM* method with the ones of the *MSPEM* method. A categorical response variable assuming $K = 3$ possible values is considered, where $k = 1$ is the reference category. Three different settings are simulated: (i) one considering only a random intercept; (ii) one considering only a random slope; (iii) one considering both random intercept and random slope[4].

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

[4] In Masci et al. (2021), the authors make this choice since in the application for modelling student dropout, the model considers a 3-categories response and only a random intercept. In the simulation study, they maintain the 3-categories response, to ease the reader, and, besides the case (i) of a random intercept, they add the other two random effects cases, in order to show the method results in more complex settings. They also include two covariates in the model (considered as both fixed or one random and one fixed) to be in line with the case study.

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

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

$$\pi_{ij1} = P(Y_{ij} = 1) = \frac{1}{1 + \sum_{l=2}^{3} \exp(\eta_{ijl})}, \tag{19}$$

where the linear predictor $\boldsymbol{\eta}_{ik} = (\eta_{i1k}, \ldots, \eta_{i200k})$ is generated in the following ways[6]:

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

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

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

(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.

All the parameters used to simulate the data and the tuning parameters of the semiparamatric method are equal to the ones used in Masci et al. (2021). In particular, we perform 100 runs of the *JMSPEM* 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 criterion , for $k = \{2, 3\}$, `tollR` $=$ `tollF` $= 0.01$, `itmax` $= 50$ and `it1` $= 30$ (see Appendix B in Masci et al. (2021) for the details). In all the runs, the *JMSPEM* algorithm converges in a number of iterations that ranges between 4 and 7, slightly quicker with respect to *MSPEM*, whose number of iterations ranges between 5 and 10. Table 1 compares the *JMSPEM* performances with the *MSPEM* ones, reporting the number of runs out of 100 in which the two methods identify the simulated number of subpopulations (i.e. $M_2 = 3$ and $M_3 = 2$) and correctly assign groups to the identified subpopulations, for all the three settings.

[6]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

JMSPEM *and* MSPEM *methods performances across the 500 runs for each of the three cases. The first two columns report the number of runs out of 500 in which the algorithms identify the correct number of subpopulations that are simulated in the data generating process (DGP) in Eq. (20), (21) and (22); third and fourth columns report the number of runs out of the number of runs in which the algorithms identify* $M_2 = 3$ *and* $M_3 = 2$ *(reported in the first two columns) in which the algorithms correctly assign each group to the correspondent subopulation.*

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

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

Table B1 reports the results of the *JMSPEM* 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 in which the estimated number of subpopulations corresponds to the simulated one (that is the majority of the cases)[7]. Estimates result to be very accurate, both for fixed and random effects, and their variability across runs is substantially low. Table B1 in Appendix reports the *MSPEM* estimated parameters. Compared to *MSPEM*, the *JMSPEM* method produces more precise and stable estimates. We observe a 93.55%, 64.12% and 51.43% decrease in the mean estimation error, for the three settings, respectively. Moreover, given that the ML estimates in multinomial regression are only *asymptotically* unbiased, we expect the performance of the algorithm to increase when the number of observations increases (Masci et al., 2021).

All details about the tuning parameters and insights about how to identify their best choice can be found in Masci et al. (2019b) and Masci et al. (2021).

[7] 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.

TABLE 2

Fixed and random effects coefficients estimated by JMSPEM *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.*

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

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

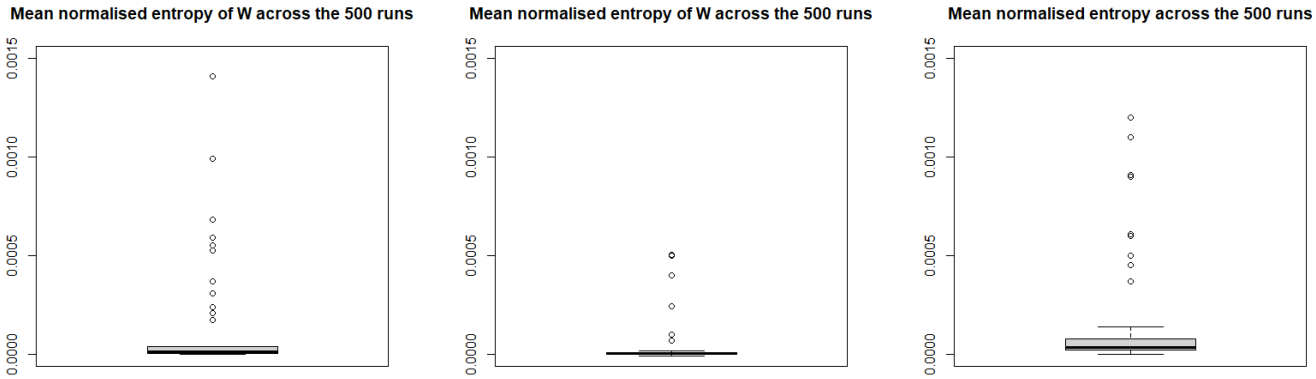
Fixed- and random effects coefficients estimated by JMSPEM *algorithm for the DGP in Eq. (21).*

	$\hat{\alpha}_k$	$\hat{b}_{1m_k k}$	$\hat{b}_{2m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_2 = -5.007 \pm 0.125$	$\hat{b}_{112} = -5.982 \pm 0.057$ $\hat{b}_{122} = -4.459 \pm 0.118$ $\hat{b}_{132} = -8.011 \pm 0.129$	$\hat{b}_{212} = 5.032 \pm 0.146$ $\hat{b}_{222} = 1.827 \pm 0.136$ $\hat{b}_{232} - 1.147 \pm 0.097$	$\hat{w}_{12} = 0.300$ $\hat{w}_{22} = 0.300$ $\hat{w}_{32} = 0.400$
	$TV = -5$	$TV = (-6, -4, -8)$	$TV = (+5, +2, -1)$	$TV = (0.3, 0.3, 0.4)$
k=3	$\hat{\alpha}_3 = 2.021 \pm 0.048$	$\hat{b}_{113} = 0.836 \pm 0.047$ $\hat{b}_{123} = -0.917 \pm 0.044$	$\hat{b}_{213} = -3.742 \pm 0.092$ $\hat{b}_{223} = 2.139 \pm 0.165$	$\hat{w}_{13} = 0.600$ $\hat{w}_{23} = 0.400$
	$TV = +2$	$TV = (+1, -1)$	$TV = (-4, +2)$	$TV = (0.6, 0.4)$

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

Extending the approach presented in Masci et al. (2021) into our multivariate setting, we evaluate the uncertainty of classification of groups into subpopulations by measuring, for each group, the normalised entropy of the conditional weights distribution. By looking at the $3-$dimensional array W, we evaluate the uncertainty of classification of each group into one of the M_{tot} $(K-1)-$variate subpopulations. Contrary to ***MSPEM***, that considers the marginal conditional weights matrices W_k, for $k = \{2, \dots, K\}$, to compute the response category-specific uncertainty of classification, by looking at the $K-$dimensional array W we compute the global uncertainty of classification of each group, with respect to all response categories. The normalised entropy of each first-dimension i of the array W is computed as the entropy $E_i = -\sum_{m=1}^{M_{tot}} W_{im} \ln(W_{im})$ divided by the maximum possible entropy value relative to M_{tot} subpopulations, i.e. $-\ln(1/M_{tot})$. We recall that the lowest level of uncertainty is reached when the algorithm assigns a group to a bivariate subpopulation m, with probability 1; in this case, the normalised entropy of the group would be equal to 0. On the opposite, the highest level of uncertainty is reached when the distribution of the conditional weights of a group i is uniform on the M_{tot} subpopulations ($W_{im} = 1/M_{tot}$ for $m_{=}1, \dots, M_{tot}$), which corresponds to an entropy $E_i = -\ln(1/M_{tot})$, and, therefore, to a normalised entropy equal to 1. The normalised entropy constitutes also a driver for the choice of the tuning parameters D_k (details in Masci et al. (2021)). Figure 1 reports the distribution of the normalised entropy of W_i, for $i = 1, \dots, I$, for the three simulated cases, mediated on the runs in which the ***JMSPEM*** algorithm identifies $M_2 = 3$ and $M_3 = 2$.

Fig 1: Boxplots of the normalised entropy of W, measured for each group, obtained by mediating the entropy in the runs in which the algorithm identifies $M_2 = 3$ and $M_3 = 2$, for the random intercept case (a), random slope case (b) and random intercept and slope case (c).



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

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

4. Case study: University student dropout across engineering degree programmes. The main novelty introduced by the *JMSPEM* algorithm is twofold. The former regards the ability to take into account and model the correlation structure across response category-specific random effects; the latter regards the positioning of the method in a tailored inferential framework.

In order to test and evaluate these aspects in a real data example, we reproduce the case study presented in Masci et al. (2021) and we compare our results with the ones obtained by both the *MSPEM* and the parametric MCMCglmm methods.

4.1. *Data and model setting.* The case study consists in the application of the method to data about Politecnico di Milano (PoliMI) students, in order to classify different profiles of engineering students and to identify subpopulations of similar degree programmes. The authors in Masci et al. (2021) consider the concluded careers of students enrolled in 19 engineering programmes of PoliMI in the academic year between 2010/2011 and 2015/2016. The dataset considers $18,604$ concluded careers of students nested within 19 engineering degree programmes (the smallest and the largest degree programmes contain 341 and $1,246$ students, respectively). 32.7% of these careers is concluded with a dropout, while the remaining 67.3% regards graduate students. The response variable regards the status of the concluded career that can be classified as:

- *graduate* - occurs when the student concludes his/her career obtaining the bachelor degree (67.3% of the sample);
- *early dropout* - occurs when the student drops within the 3^{rd} semester after the enrolment (16.2% of the sample);
- *late dropout* - occurs when the student drops after the 3^{rd} semester after the enrolment (16.5% of the sample).

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

Regarding student characteristics, besides the status of the concluded career and the degree programme the student is enrolled in, the number of European Credit Transfer System credits (ECTS), i.e. the credits he/she obtained at the first semester of the first year of career (the variable has been standardized in order to have 0 mean and 1 sd) and his/her gender (the sample contains 22.3% females and 77.7% males) are considered[8]. Table 3 reports the variables considered in the analysis with their explanation. For further information on the original dataset and the preprocessing phase please refer to Masci et al. (2021).

[8]In Masci et al. (2021), the authors state that only information at the first semester of career is used because it is observable for all students (either dropout or graduate) and it allows to predict student dropout as soon as possible, i.e. at the beginning of the student career.

TABLE 3
List and explanation of variables at student level to be included in the model (Masci et al., 2021)

Variable	Description	Type of variable
`Status`	Status of concluded career	3-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

The modelling proposed is the following. 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 form:

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

where

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

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 $2-$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 programme (`DegProg`) and to the $k-$th category.

Given the data setting and model formulation presented in Eq.s (23, 24, 25) , we apply the *JMSPEM* algorithm to PoliMI data and we compare the results with the ones obtained by applying the *MSPEM* algorithm and the parametric MCMCglmm method. The aim of the study is to model the probability of being *early* or *late dropout* student, with respect to being a *graduate* one, given student characteristics and early career information, and considering the nested structure of students within the 19 degree programmes. Both *MSPEM* and *JMSPEM* algorithms, by assuming discrete random effects, identify subpopulations of degree programmes, depending on their effects on *early* and *late dropout* probability, while the MCMCglmm algorithm, by assuming Gaussian random effects, identifies a ranking of degree programmes.

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

4.2. *JMSPEM results.* We run the *JMSPEM* algorithm with the same parameters setting chosen in Masci et al. (2021): `tollR`=`tollF`=10^{-2}, `itmax`=60, `it1`=20, $\tilde{w} = 0$ and $D_k = 0.3$, for $k = 2, 3$. The algorithm converges in 9 iterations and identifies 5 supopulations for both categories $k = 2$ (early dropout) and $k = 3$ (late dropout). Tables 4 and 5 report the estimated model parameters and the distributions of the 19 degree programmes across the identified subpopulations, respectively.

TABLE 4

Fixed and random effects coefficients estimated by JMSPEM *algorithm for student dropout prediction. Standard errors of the estimates are reported in brackets. Asterisks denote different levels of significance: . 0.01 < p-val < 0.1; * 0.001 < p-val < 0.01; ** 0.0001 < p-val < 0.001; *** p-val < 0.0001.*

	$\hat{\alpha}_{1k}$ (`Gender`)	$\hat{\alpha}_{2k}$ (`TotalCredits1.1`)	$\hat{b}_{m_k k}$ (random intercept `DegProg`)	$\hat{w}_{m_k k}$ (weight)
k=2	$\hat{\alpha}_{12} = 0.014(0.0609)$	$\hat{\alpha}_{22} = -2.684^{***}(0.0218)$	$\hat{b}_{12} = -3.504(0.0780)$	$\hat{w}_{12} = 0.100$
			$\hat{b}_{22} = -3.023(0.0710)$	$\hat{w}_{22} = 0.167$
			$\hat{b}_{32} = -2.485(0.0385)$	$\hat{w}_{32} = 0.291$
			$\hat{b}_{42} = -2.138(0.0537)$	$\hat{w}_{42} = 0.391$
			$\hat{b}_{52} = -1.728(0.0429)$	$\hat{w}_{42} = 0.051$
k=3	$\hat{\alpha}_{13} = -0.577^{***}(0.0606)$	$\hat{\alpha}_{23} = -1.907^{***}(0.0211)$	$\hat{b}_{13} = -2.491(0.1001)$	$\hat{w}_{13} = 0.147$
			$\hat{b}_{23} = -1.950(0.0566)$	$\hat{w}_{23} = 0.173$
			$\hat{b}_{33} = -1.601(0.0386)$	$\hat{w}_{33} = 0.321$
			$\hat{b}_{43} = -1.245(0.0519)$	$\hat{w}_{43} = 0.144$
			$\hat{b}_{53} = -0.903(0.0430)$	$\hat{w}_{53} = 0.215$

TABLE 5

Distribution of the 19 degree programmes across the 5 identified subpopultions relative to $k = 2, 3$. For each k, subpopulations are ordered from 1 to 5 coherently with the estimated coefficients reported in Table 4.

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

Focusing on the random effects, last two columns in Table 4 report, for each k, the random intercepts associated to the five subpopulations with their weights, ordered increasingly. The *JMSPEM* algorithm identifies Biomedical Engineering as the degree programme in which students are more likely to early drop, all else equal, while Civil and Environmental engineering and Environmental and Land Planning engineering result to be the ones in which students tend to early drop less than the others, all else equal (Table 5). These two subpopulations have relatively lower weights with respect to the other three subpopulations, that represent the majority of the sample, and, consequently, are interpreted as the ones containing three degree programmes with anomalous behaviors. For late dropout, degree programmes are more uniformly distributed across the five subpopulations, starting from Subpopulation 1, that contains the three degree programmes associated to the lowest late dropout probability, until Subpopulation 5, that contains the four degree programmes associated to the highest late dropout probability.

We evaluate the uncertainty of classification by measuring, for each degree programme $i = 1, \ldots, 19$, the normalised entropy of the conditional weights, computed as $E_i = -\sum_{m=1}^{M_{tot}} W_{im} \ln(W_{im})$ divided by the maximum possible entropy value relative to M_{tot} subpopulations, i.e. $-\ln(1/M_{tot})$, where $M_{tot} = 5 \times 5 = 25$. Mean and median of the 19 normalised entropy distribution are 0.0785 and 0.0426, respectively, while minimum and maximimum values are 0.0002 and 0.2681,respectively, indicating a low level of uncertainty of classification.

Besides the support points and relative weights of the two marginal discrete distributions of random-effects B_2 and B_3 reported in Table 4, we estimate their variance-covariance matrix, the correlation between B_2 and B_3 and the VPCs. Given the estimated support points $b_{m_k k}$, for $m_k = 1, \ldots, M_k$ and $k = \{2, 3\}$, and relative weights, the variance σ^2_{rk} of the two marginal distributions of random effects can be computed, thanks to the Eve's law, as

$$\sigma^2_{rk} = Var[B_k] = E[Var[B_k|(b_{1k}, \ldots, b_{M_k k})]] + Var[E[B_k|(b_{1k}, \ldots, b_{M_k k})]], \tag{26}$$

where

$$E[Var[B_k|(b_{1k}, \ldots, b_{M_k k})]] = \\ = E[B_k^2|(b_{1k}, \ldots, b_{M_k k})] - (E[B_k|(b_{1k}, \ldots, b_{M_k k})])^2 = \\ = \sum_{m_k=1}^{M_k} b^2_{m_k k} \times w_{m_k k} - \left(\sum_{m_k=1}^{M_k} b_{m_k k} \times w_{m_k k} \right)^2$$

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

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

For $k = \{2, 3\}$, by summing up these two quantities in Eq. (26), we obtain

$$\sigma^2_{r2} = 0.2275 + 0.0146 = 0.2421$$

and

$$\sigma^2_{r3} = 0.2603 + 0.0006 = 0.2609.$$

In order to compute the covariance, we refer to the estimated 5×5-matrix of joint weights $\mathbf{w}$ [9]:

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

Thanks to this quantity, we can compute the covariance between B_2 and B_3 as:

$$\begin{aligned} Cov(B_2, B_3) &= E[B_2 B_3] - E[B_2]E[B_3] = \\ &= \sum_{m=1}^{M_2 \times M_3} w_m \times b_{m2} \times b_{m3} - \left(\sum_{m_2=1}^{M_2} w_{m_2 2} \times b_{m_2 2} \right) \times \left(\sum_{m_3=1}^{M_3} w_{m_3 3} \times b_{m_3 3} \right) = \\ &= 4.1553 - (-2.5024) \times (-1.5916) = 0.1725. \end{aligned}$$

The variance-covariance matrix of $\mathbf{B}$ is, therefore,

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

and the correlation between B_1 and B_2 is 0.6863, that is in line with what we expected by looking at Panel (a) in Figure 2.

To measure the significance of the random-effects estimated by *JMSPEM*, we compute the VPC relative to each logit. For each $k = \{2, 3\}$, the portion of the total variability in the response explained by the latent structure identified at the degree programmes level is

$$VPC_k = \frac{\sigma^2_{rk}}{\sigma^2_{rk} \times \pi^2/3},$$

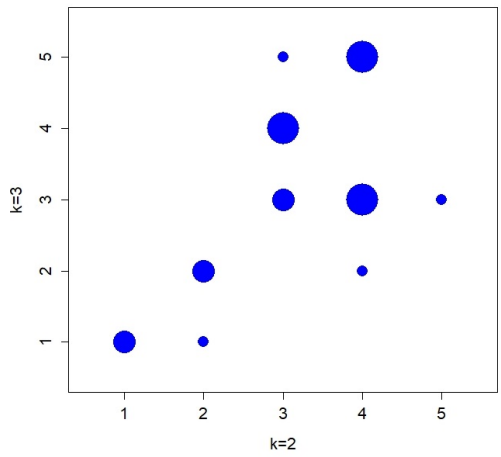
that corresponds to $VPC_2 = 0.06857$ and $VPC_3 = 0.07348$, respectively. For both early and late dropout, about 7% of the total variability is explained by the subpopulations structure. Results of MCMCglmm provide $VPC_2 = 0.0906$ and $VPC_3 = 0.1091$.

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

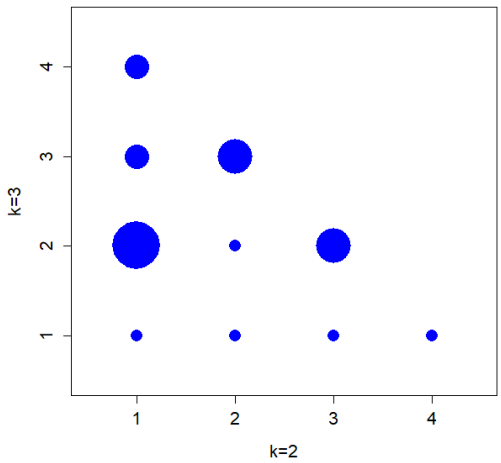
4.3. *Comparison between* JMSPEM*,* MSPEM *and MCMCglmm.* For comparing *JMSPEM* results with the ones obtained by applying *MSPEM* and MCMCglmm algorithms to this case study, we report in Appendix fixed and random effects estimates of the two alternative methods. Starting from fixed effects, Tables A1 and A2 in Appendix report the fixed effects coefficients estimated by *MSPEM* and MCMCglmm, while first two columns in Table 4 report *JMSPEM* fixed-effects estimates. Results of the three methods are comparable: *MSPEM* and *JMSPEM* estimated coefficients are very close to each other and coherent with the MCMCglmm ones. Furthermore, the significant coefficients estimated by *JMSPEM* and MCMCglmm are the same. *MSPEM* algorithm does not include any measurement of standard errors or coefficients significance. In particular, both *JMSPEM* and MCMCglmm show that females have, on average, a lower probability of late dropout with respect to males, while no significant gender difference emerges for early dropout, and that the number of credits obtained at the first semester is inversely proportional to the probability of both early and late dropout. Fixed effects result to be robust and invariant with respect to different random effect assumptions.

Regarding the random effects, Tables A1 and A3 in Appendix report the *MSPEM* estimates. Comparing the subpopulations identified by the two semi-parametric methods we notice some differences. Both *MSPEM* and *JMSPEM* identify Biomedical engineering as the degree programme in which students are more likely to early drop. For late dropout, both *MSPEM* and *JMSPEM* algorithms assign Civil and Environmental engineering and Environmental and Land Planning engineering to the subpopulation associated to the highest late dropout probability. The remaining of the distributions of degree programmes on the estimated subpopulations is more heterogeneous across the two methods. What is also interesting to compare between *MSPEM* and *JMSPEM* results is the distribution of degree programmes on the bivariate subpopulations, displayed in Figure 2. Each bubble size is proportional to the weight of the bivariate subpopulation. From an interpretative point of view, Figure 2 helps us in comparing the effect of each degree programmes with respect to early and late dropout. For *JMSPEM*, the distribution of the weights on the bisector of the figure in Panel (a) suggests that, except for very few cases (e.g., Biomedical Engineering), degree programmes effects are quite coherent between early and late dropout. On the opposite, the distribution of the weights of the bivariate subpopulations obtained by applying the *MSPEM* algorithm (Panel (b)) suggests that degree programmes in which students are more likely to early drop are less likely to late drop and vice-versa. This result demonstrates that different assumptions on the dependence structure across random effects distributions lead to relevant differences in the estimates and in their interpretation.

Fig 2: Distribution of the weights of the *JMSPEM* (panel (a)) and *MSPEM* (panel (b)) estimated bivariate subpopulations. Bubble size is proportional to the number of degree programmes belonging to the subpopulations couple. Image in panel (b) is taken from Masci et al. (2021).



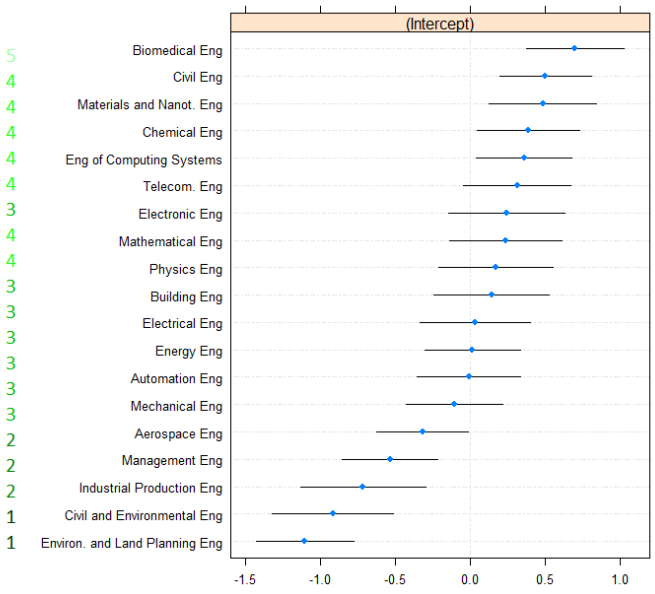
(a) *JMSPEM* degree programmes distribution across subpopulations.



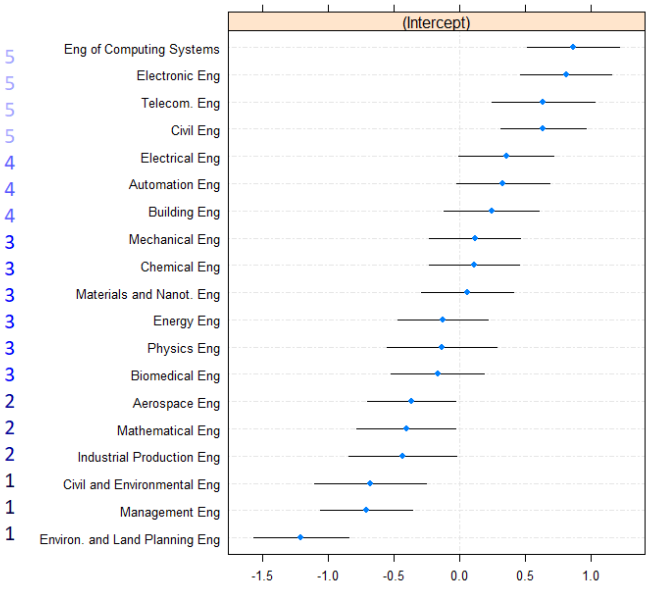
(b) *MSPEM* degree programmes distribution across subpopulations.

Regarding the comparison with the parametric MCMCglmm approach, Figure 3 shows the rankings of the estimated random intercepts, with their confidence intervals, for early (Panel (a)) and late dropout (Panel (b)), respectively. In order to ease the comparison with *JMSPEM* results, we annotated the subpopultion number, estimated by *JMSPEM*, alongside the degree programmes names. The *JMSPEM* subpopulations and the MCMCglmm rankings are extremely coherent. The same does not holds for the *MSPEM* results. Figure A1 in Appendix shows the same MCMCglmm ranking, alongside which are reported the Subpopulations estimated by *MSPEM* algorithm (Masci et al., 2021): the matching between the ranking and the subpopulations is less precise.

Fig 3: Panels (a) and (b) show the ranking of 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. Alongside degree programmes names, we report subpopulations indexes estimated by *JMSPEM* algorithm. Colours are only intended to help in the visualization.



(a) Degree programmes intercepts for Early Dropout



(b) Degree programmes intercepts for Late Dropout

Lastly, we evaluate and compare the goodness of fit of the three methods, by computing their relative misclassification tables (Table 6).

TABLE 6
Misclassification tables relative to JMSPEM *(left tabular),* MSPEM *(central tabular) and* MCMCglmm *(right tabular) predictions, expressed in percentages.*

	obs D1	obs D2	obs G
pred D1	0.099	0.060	0.018
pred D2	0.033	0.043	0.014
pred G	0.032	0.064	0.637

	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

Error rates are 22.1% for *JMSPEM*, 21.6% for MCMCglmm and 23.3% for *MSPEM*, respectively. As noted in Masci et al. (2021), we expect the MCMCglmm to have the best fit, since it estimates a single random effect for each degree programme (and, therefore, it fits the data 'deeply'). *JMSPEM* error rate is lower than the *MSPEM* one and it is very close to the MCMCglmm one, suggesting that the identified subpopulations catch almost the entire heterogeneity across degree programmes effects. This is somehow expected since *JMSPEM* and MCMCglmm modellistic assumptions are more flexible and less strict with respect to *MSPEM* ones, leading to a better capacity to model the real dynamics within the data.

Given the high predictive performance and the matching with the parametric approach, the *JMSPEM* algorithm proves to produce precise and reliable estimates.

5. Concluding remarks and future perspectives. In this paper, we propose a mixed-effects model with discrete random effects for an unordered multinomial response, together with a suitable inferential framework. Estimates of parameters are obtained through an Expectation-Maximization algorithm, called *JMSPEM*. The proposed method is an extension of the *MSPEM* algorithm presented in Masci et al. (2021), in which we relax the independence assumption across response categories. The *JMSPEM* algorithm consists in a semi-parametric approach that assumes the response category-specific random effects to follow a discrete distribution with an *a priori* unknown number of mass points, that are allowed to differ across response categories. With respect to the traditional parametric approach, the *JMSPEM* algorithm constitutes a valid alternative, both from a computational and an interpretative point of view. Indeed, the discrete distribution on the random effects allows to write the likelihood function as a weighted sum, avoiding integration issues typical of parametric mixed-effects multinomial models, and, moreover, allows to identify a latent structure of subpopulations at the highest level of grouping. With respect to the existing *MSPEM* algorithm, that has been developed under the independence assumption across the response category-specific random effects distributions, the *JMSPEM* algorithm, by relaxing this often too strict and unrealistic assumption, results to be a more sophisticated and flexible method. Besides its potential to take into account and model more complex data structures, the *JMSPEM* algorithm produces more accurate estimates and provides a measure of the significance and the uncertainty of the estimates.

After describing the *JMSPEM* method, we reproduce the simulation study and the case study reported in Masci et al. (2021), in order to test and evaluate the performances of the *JMSPEM* algorithm, compared to the *MSPEM* ones. In confirmation of what is expected from a theoretical point of view, taking into account the dependence structure that is naturally intrinsic within the data results to be a significant value added. Results of the simulations show that *JMSPEM* produces very accurate estimates, with a reduced bias with respect to the *MSPEM* estimates. The *JMSPEM* fitting and predictive power is also confirmed when applied to the real data example. In the context of predicting the types of concluded careers of Politecnico di Milano students, nested within different engineering degree programmes, the *JMSPEM* algorithm proves higher predictive performance compared to *MSPEM*. Moreover, the estimated subpopulations of degree programmes, that differ from the ones estimated by *MSPEM*, are extremely coherent with the ranking obtained by applying the parametric MCMCglmm, proving the relevant effect that the assumption on the random effects dependence structure has on the results.

This paper enters both in the literature about multinomial regression (Agresti, 2018) and in the one about mixed-effects models with discrete random effects (Aitkin, 1999; Hartzel, 2000; Masci et al., 2019b). The proposed method contributes to both the streams but, at the same time, suffers from some of their typical criticalities. Given the presence of multiple logits, multinomial regression models are often treated as multivariate models and, in addition, the likelihood function is such that its maximization in closed form is not feasible. These two aspects contribute to require an important computing power and numerical methods for the maximization steps. For what concerns mixed-effects models with discrete random effects, we believe that they are extremely useful in many different contexts of application and that the research of a latent structure of subpopulations at the highest level of grouping is an innovative and interesting way of analysing this level of the hierarchy. Their application to real data in which the cardinality of the groups is very high and in which subpopulations are *a posteriori* explained can be extremely informative. Nonetheless, although these methods do not require to fix the number of subpopulations *a priori* but they estimate it together with the other parameters, this estimate is extremely sensitive to the choice of the threshold distance D. Some criteria to choose D have been proposed in the literature (Masci et al., 2019b, 2021)

but its choice is still sensitive and impacting. For these reasons, future work will be devoted to the embedding of more efficient optimization algorithms and to the development of a clear rule to drive the choice of the threshold distance D.

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

APPENDIX A: MSPEM AND MCMCGLMM ALGORITHMS RESULTS FOR THE POLIMI CASE STUDY

This section reports the results of the *MSPEM* algorithm and of the MCMCglmm method applied to PoliMI data (Masci et al., 2021). In particular, Table A1 reports the fixed and random effects coefficients estimated by *MSPEM* algorithm for student dropout prediction; Table A2 reports fixed effects coefficients estimated by MCMCglmm algorithm for student dropout prediction.

TABLE A1
Fixed and random effects coefficients estimated by MSPEM *algorithm for student dropout prediction.*

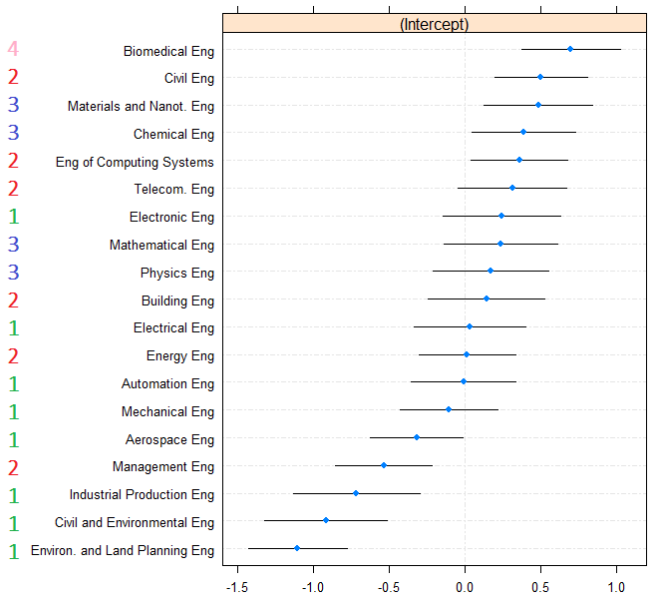
	$\hat{\alpha}_{1k}$ (`Gender`)	$\hat{\alpha}_{2k}$ (`TotalCredits1.1`)	$\hat{b}_{m_k k}$ (random intercept `DegProg`)	$\hat{w}_{m_k k}$ (weight)
k=2	$\hat{\alpha}_{12} = -0.153$	$\hat{\alpha}_{22} = -2.704$	$\hat{b}_{12} = -2.841$ $\hat{b}_{22} = -2.423$ $\hat{b}_{32} = -2.096$ $\hat{b}_{42} = -1.586$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

TABLE A2
Fixed effects coefficients estimated by MCMCglmm algorithm for student dropout prediction.

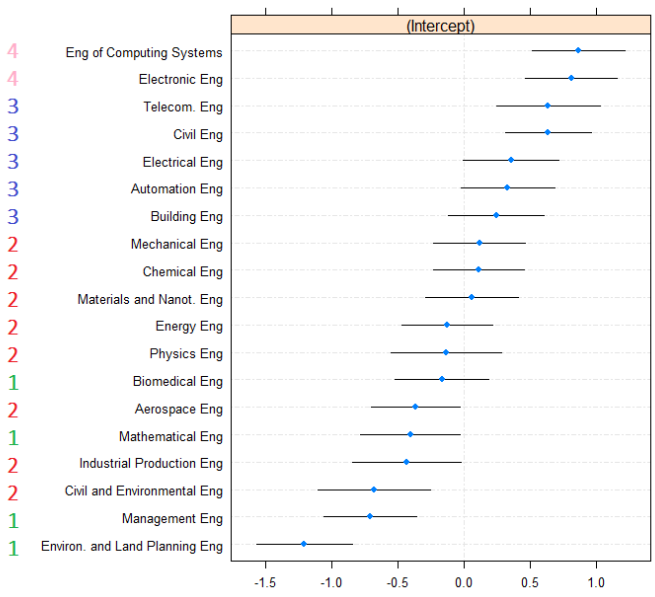
	Variable name	post.mean	$l-95\%$ CI	$u-95\%$ CI	pMCMC
k=2	`Intercept`	-2.552	-2.854	-2.269	$< 0.001 **$
	`Gender`	-0.027	-0.106	0.153	0.769
	`TotalCredits1.1`	-2.797	-2.884	-2.702	$< 0.001 **$
k=3	`Intercept`	-2.354	-2.672	-2.049	$< 0.001 **$
	`Gender`	-0.634	-0.464	-0.802	$< 0.001 **$
	`TotalCredits1.1`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

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

Fig A1: 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. Alongside the degree programmes names, subpopulations indexes estimated by *MSPEM* algorithm are reported (Masci et al., 2021). Colours are only intended to help in the visualization.



(a) Degree programmes intercepts for Early Dropout



(b) Degree programmes intercepts for Late Dropout

TABLE A3

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 A1.

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		

APPENDIX B: MSPEM RESULTS OF THE SIMULATION STUDY

TABLE B1

Table of fixed and random effects coefficients estimated by MSPEM *algorithm in the simulation study, first presented in Masci et al. (2021) and reproduced in this paper. 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 $M_3 = 2$ (shown in Table 1 in Masci et al. (2021) and here 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.*

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

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

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

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

	$\hat{\alpha}_k$	$\hat{b}_{1m_k k}$	$\hat{b}_{2m_k k}$	$\hat{w}_{m_k k}$
k=2	$\hat{\alpha}_2 = -5.013 \pm 0.098$	$\hat{b}_{112} = -5.863 \pm 0.236$ $\hat{b}_{122} = -4.700 \pm 0.129$ $\hat{b}_{132} = -8.022 \pm 0.237$	$\hat{b}_{212} = 5.091 \pm 0.195$ $\hat{b}_{222} = 2.801 \pm 0.119$ $\hat{b}_{232} = -1.185 \pm 0.079$	$\hat{w}_{12} = 0.300$ $\hat{w}_{22} = 0.300$ $\hat{w}_{32} = 0.400$
	$TV = -5$	$TV = (-6, -4, -8)$	$TV = (+5, +2, -1)$	$TV = (0.3, 0.3, 0.4)$
k=3	$\hat{\alpha}_3 = 1.977 \pm 0.040$	$\hat{b}_{113} = 0.739 \pm 0.058$ $\hat{b}_{123} = -0.888 \pm 0.055$	$\hat{b}_{213} = -3.651 \pm 0.092$ $\hat{b}_{223} = 2.419 \pm 0.160$	$\hat{w}_{13} = 0.600$ $\hat{w}_{23} = 0.400$
	$TV = +2$	$TV = (+1, -1)$	$TV = (-4, +2)$	$TV = (0.6, 0.4)$

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

REFERENCES

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

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

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.

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

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.

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.

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

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

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.

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

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

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.

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

Maggioni, A. (2020). *Semi-parametric generalized linear mixed effects model: an application to engineering BSc dropout analysis*. PhD thesis.

Masci, C., Ieva, F., Agasisti, T., and Paganoni, A. M. (2019a). Evaluating class and school effects on the joint achievements in different subjects: a bivariate semi-parametric mixed-effects model. *MOX-report n^o 24/2019*.

Masci, C., Ieva, F., and Paganoni, A. M. (2021). Semiparaemtric multinomial mixed-effects models: a university students profiling tool. *The Annals of Applied Statistics*, in press.

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

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

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. https://www.R-project.org/.

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

Raudenbush, S. W., Yang, M.-L., 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.

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. *British Journal of Mathematical and Statistical Psychology*, 69(3):316–343.

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.

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.

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

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