ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2019)

Semiparametric mixed effects models for unsupervised classification of Italian schools

Chiara Masci, Anna Maria Paganoni and Francesca Ieva

Politecnico di Milano, Italy

[Received November 2017. Final revision February 2019]

Summary. The main purpose of the paper is to improve research on school effectiveness by applying a new strategy for uncovering subpopulations of schools that differ in terms of distribution of student outcomes. We propose a semiparametric mixed effects model with an expectation–maximization algorithm to estimate its parameters and we apply it to the Italian Institute for the Educational Evaluation of Instruction and Training data of 2013–2014 as a tool for the identification of latent subpopulations of schools. The semiparametric assumption provides the random effects of the mixed effects model to be distributed according to a discrete distribution with an (*a priori*) unknown number of support points. This modelling induces an automatic clustering of schools (the higher level of hierarchy), where schools within the same cluster share the same random effects. The latent subpopulations of schools identified may then be exploited through the use of multinomial models that include school level features. The novelties introduced by this paper are twofold: first, the semiparametric expectation–maximization algorithm is an innovative method that could be used in many classification problems; second, its application to education data represents a new approach to study school effectiveness.

Keywords: Expectation–maximization algorithm; School value added; Semiparametric mixed effects models; Student achievements

1. Introduction

The analysis of education systems is a subject that has received particular attention in recent decades. During their learning process, students are influenced by multiple aspects coming from both their personal life and their school life. Personal motivation, family, friends and geographical context play a fundamental role in students' performance and choice of school is also particularly relevant. The literature provides numerous studies aimed at measuring and explaining the 'school effect', intended as the influence that the school the student is attending has on his or her achievements with respect to other schools; see, among others, Bryk and Raudenbush (1988), Coleman *et al*. (1966), Hanushek *et al*. (1996) and Raudenbush and Bryk (1986). Bryk and Raudenbush (1988) stated the importance of considering the 'unit of analysis' (students, classes and schools), when speaking about educational research, and they argued that hierarchical models should constitute the basic paradigm for quantitative research on student learning. Also, given the hierarchical structure of education data, Raudenbush and Bryk (1988) underlined the importance of measuring school effects, as intended before, and presented various approaches to analyse nested data. Coleman *et al*. (1966) viewed education as a process in which students' performance (output) is produced from inputs including school resources, teacher

Address for correspondence: Chiara Masci, Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, via Bonardi 9, Milan, Italy.
E-mail: chiara.masci@polimi.it

 0964–1998/19/182000

quality, family attributes and peer quality. In their perspective, policy should be focused on inputs that are both directly controlled by policy makers (characteristics of schools, teachers, curricula, etc.) and those that are 'uncontrolled' (family, friends, the learning capacities of the student, etc.). Also Coleman *et al.* (1966) showed that school characteristics are of importance in determining student outcomes.

The nature and the magnitude of the school impact on the attainments of students strongly depend on the type of school system and related regulations. There are countries where the education system is totally centralized and, therefore, school programmes and practices are very homogeneous across the territory. In contrast, in recent years the dynamics of education systems have been changing and increasingly more countries are decentralizing the power for deciding about education, giving more autonomy to schools (Sarrico *et al.*, 2012). This leads to differences between schools that are reflected in differences between students' achievements. The Organisation for Economic Co-operation and Development's Programme for International Student Assessment (`www.pisa.oecd.org`) has tested 15-year-old students in mathematics, reading and science in more than 70 countries all over the world, every 3 years since 2000. Studies on data from the programmes show that Italy is a country where the percentage of variability in student achievements due to the grouping factor (i.e. schools) is quite high with respect to other countries (Masci *et al.*, 2018). This means that in Italy the value added, which is seen as the positive or negative effect, that schools give to their students is relevant: in other words, attending a certain school instead of another might lead to different results in students' skills. Schools differ in many respects: size, location, school body composition, teachers, school principal management style etc. All these aspects contribute to the students' learning process, creating heterogeneity within their achievements.

Focusing on the Italian context, the National Institute for the Educational Evaluation of Instruction and Training (the Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione, known as 'INVALSI') has tested students all over the country since 2004, in both their mathematics and their literature skills, following a procedure similar to the Organisation for Economic Co-operation and Development's Programme for International School Assessment. These tests are administered at several grades, starting from primary schools up to the end of secondary schools, producing longitudinal data that collect multiple observations for each student. Also, at national level, many studies have confirmed that the magnitude of the school effect, intended as the positive or negative value added of the school, on student attainments is substantial. Agasisti *et al.* (2017) and Masci *et al.* (2016b, 2017) observed that the percentage of variability in student attainments in INVALSI tests explained by the random effect depends on the geographical macroarea and differs between mathematics and reading performances. In particular, this percentage is higher in mathematics and especially in southern Italy, reaching peaks of 20%. Moreover, results from the Programme for International Student Assessment data in Italy report that, in mathematics, the percentage of variability explained by random effects, PVRE, exceeds 40% (Masci *et al.*, 2019).

An important characteristic of educational administrative data is their hierarchical structure: students are naturally nested within schools. In the learning process investigation, it is important to disentangle the effects that are given by each level of hierarchy and, to the best of our knowledge, multilevel models are one of the best tools to fit the nature of nested data (see, for example, Bock (2014) and Vanthienen and De Witte (2017)). Indeed, multilevel models take into account the hierarchical nature of data and can quantify the part of variability in the response variable that is given to each level of grouping (Pinheiro and Bates, 2000). In particular, in the case of students nested within schools, they could estimate the 'school effect', i.e. the value added (i.e. the positive or negative effect) of the school to its student achievements.

In this perspective, the aim of this work is to identify latent subpopulations of Italian schools that differ in the evolution of their student attainments across years. The goal is to reduce the set of numerous Italian schools into a series of subpopulations, each of which contains schools with a similar effect on student achievements and across which these impacts differ. For this, we need a model that takes into consideration the hierarchical structure of data, but that also identifies a latent structure among the higher level of hierarchy. Therefore, we apply a multilevel model in which we model the subpopulations by choosing as random effects a discrete distribution P^* with an unknown finite number of mass points that can detect a latent structure among the Italian schools: the higher level of hierarchy. This model can be interpreted as an in-built unsupervised classification tool, since it identifies a clustering structure among groups, without knowing *a priori* either the number of clusters or the number of groups per cluster. From a practical point of view, in Italy students must attend 5 years of primary school, 3 years of junior secondary school and 5 years of upper secondary school. We are aware of the challenges that estimating the pure school effect implies (Goldstein and Spiegelhalter, 1996; Raudenbush and Willms, 1986); indeed, we shall not refer to a school effect in the classical way. Rather, since we focus on junior secondary schools, our school effect can be interpreted as the ability of these schools in receiving students from primary schools with certain skills to give them new and possibly increased skills at the end of the 3 years, aware of the fact that students might not be randomly assigned to schools. So our research mainly aims at identifying subpopulations of schools, with respect to the relationship between their students' test scores at the beginning and at the end of the 3 years (grades 6 and 8 respectively). Supposing that we can model the relationship between students' test scores at different grades by means of linear models, which means that student scores at different grades are assumed to be linearly correlated, the regression line between the two grades test scores might be characterized by different parameters across schools. In other words, we try to identify subpopulations of Italian junior secondary schools, characterized by different trends in their student achievements, where the number of subpopulations is unknown *a priori*.

In the methodological literature, two lines of research about the identification of subpopulations are

(a) growth mixture models (GMMs) and
(b) latent class mixture models (McCulloch *et al.*, 2002; Nagin, 1999; Vermunt and Magidson, 2002).

Conventional growth modelling is applied to longitudinal data and is used to estimate the average growth, the amount of variation across individuals in growth intercept and slopes and the influence of covariates on this variation. It can be described as a random-effect model where the intercept and slope vary across individuals. However, conventional growth models assume that individuals come from a single population and that a single growth trajectory can approximate the entire population. GMMs relax this assumption and assume that there are differences in growth parameters across unobserved subpopulations. They allow for the existence of latent trajectory classes where different groups of individual growth trajectories vary around different behaviours. In other words, the average association between covariates and the outcome varies across latent classes and also, within classes, individuals vary randomly in their coefficients. The results are separate growth models for each latent class. Latent class growth analysis (LCGA) is a special case of GMMs where the variance and covariance parameters are assumed to be 0, implying that all the individuals within a latent class are homogeneous. Individuals within a latent class are assumed to have identical random effects. Conceptually, these methods are very similar to the method that we propose, especially the special case of LCGA, since we also

assume that individuals within latent classes have identical random effects. Nonetheless, there are two main differences between our approach and those of GMMs and LCGA. First GMMs or LCGA are tailored to model longitudinal changes and not variation within groups. (One of the characteristics of the models for longitudinal data is that the set of time instants in which the dependent variable is evaluated is the same within each group or individual, meaning that the covariate is fixed across the groups or individuals.) Second GMMs or LCGA need to fix *a priori* the number of latent classes, whereas our approach estimates it together with the other unknown parameters. There are numerous extensions and applications of GMMs (Lin, 2000; Proust-Lima *et al*. 2007), but none of them includes the estimation of the number of latent classes. Indeed, these past methods require that the analysts estimate a series of models, where each model assumes a different number of clusters, and then use model fit statistics to compare these models to select the best fitting model. In our approach, the analyst specifies a caliper, the maximum distance between two clusters such that the two can be collapsed, and the algorithm then estimates the number of clusters. Latent class mixture models are even more related to our approach since they consider linear mixed models where the assumption of normality of random effects is relaxed. They also assume a discrete distribution for the random-effect coefficients and they are used to uncover distinct subpopulations (latent classes) and to classify individuals. But also this approach requires a fixed number of latent classes, chosen *a priori*. In the framework of latent structure analysis, another branch of research that is related to ours is latent trait analysis (Bock and Aitkin, 1981; Heinen, 1996). Latent trait analysis, which is also called item response theory, is used for the analysis of categorical data. It performs the reduction of a set of binary or ordered categorical variables into a smaller set of factors and it is used both to calibrate items and to derive latent trait estimates that are then used in subsequent analysis. The common aspect of this method with those described above and, at the same time, the main difference from our method is, again, the fact that they need to fix *a priori* the number of latent classes. The choice of the number of latent classes (mass points) is not trivial when the sample is very big or knowledge about possible different trends across the individuals (groups) is limited. Our case-study represents a clear example of a sample composed of hundreds of groups, within which we do not know how many different subpopulations exist. For this reason, in performing dimensionality reduction without any assumption about the final dimension, we need to develop an approach that estimates, together with the other parameters, also the number of existing subpopulations. In this sense, our approach brings significant value added with respect to the existing literature.

In particular, we develop and apply an expectation–maximization (EM) algorithm for semiparametric mixed effects models (Bock and Aitkin, 1981) for hierarchical data (students nested within schools), to perform an in-built classifier of the grouping factor (schools). The algorithm is inspired by those proposed in Aitkin (1996) and Azzimonti *et al*. (2013), but with substantial changes. The idea is that we perform a linear two-level model, in which we consider students nested within schools, where the random effect (school effect) is semiparametric since it follows a discrete distribution with an unknown number of support points. The algorithm itself identifies the number of support points, i.e. the number of subpopulations in which schools are grouped, with respect to the trend of achievements of their students. In the educational literature, multilevel linear models have already been applied to INVALSI data, with a view to estimating school value added, modelled by means of parametric distributions, after adjusting for student characteristics (Agasisti *et al*., 2017; Masci *et al*., 2016, 2017; Sani and Grilli, 2011). Nonetheless, our method has a different scope since it does not seek to estimate individual value added for each school, but it looks for subpopulations of schools with homogeneous value added. Both the algorithm and its application to the educational context are new to the literature.

From an interpretative point of view, the identification of subpopulations of schools reveals how many and which different behaviours characterize Italian schools and, therefore, identifies a latent structure within them. In particular, the distribution of schools across subpopulations tells us which is the most numerous subpopulation and identifies subpopulations of anomalous schools, i.e. those subpopulations containing fewer schools with different impact on student achievements. Once schools have been classified into clusters in a second stage we aim at profiling the subpopulations by means of school level variables, analysing which school level characteristics predict cluster membership. The idea is that there could be variables at school level that influence the different student achievement trends across schools. Therefore, in the second part of the analysis we explore the presence of patterns of school characteristics among subpopulations of schools by means of multinomial regression models.

The paper is organized as follows: in Section 2 we describe the model and methods—the semiparametric EM algorithm—and we present a simulation study; in Section 3 we describe the INVALSI data set and report the application of the semiparametric EM algorithm to INVALSI data, show the results and explore the relationship between subpopulations and school characteristics; in Section 4 we draw our conclusions.

All the analysis are made by using R software (R Development Core Team, 2014). The code for the semiparametric EM algorithm is available on request to the authors.

2. Model, methods and simulation study

In this section, we present the semiparametric mixed effects model (Section 2.1), the EM algorithm for the estimation of its parameters (Section 2.2) and a simulation study (Section 2.3). Since we know from previous research on Italian data that there are patterns of student achievements across different Italian schools (Agasisti *et al.*, 2017; Masci *et al.*, 2016, 2017), we are interested in evaluating how the association between previous and current student test scores changes across different Italian schools and, in particular, in identifying subpopulations of schools within which this association is identical. Therefore, the model that we develop is a two-level linear model (in the application, students represent level 1 and schools represent level 2) with a discrete distribution with a finite number of support points on the random effects. This modelling enables us to identify a latent structure of subpopulations in the higher level of grouping (in the application, schools).

2.1. Semiparametric mixed effects model

Consider a general mixed effects (two-level) linear model, where each observation j, for $j = 1, \ldots, n_i$, is nested within a group i, for $i = 1, \ldots, N$. The model takes the form

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N, \\ \boldsymbol{\epsilon}_i &\overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{1}_{n_i}) \end{aligned} \tag{1}$$

where i is the group index, N is the total number of groups, n_i is the number of observations within the ith group and $\Sigma_{i=1}^{N} n_i = J$. $\mathbf{y}_i = (y_{1i}, \ldots, y_{n_i i})$ is the n_i-dimensional vector of response variable within the ith group, $\mathbf{X}_i$ is the $n_i \times (p+1)$ matrix of covariates having fixed effects, $\boldsymbol{\beta}$ is the $(p+1)$-dimensional vector of fixed coefficients, $\mathbf{Z}_i$ is the $n_i \times (r+1)$ matrix of covariates having random effects, $\mathbf{b}$ is the $(r+1)$-dimensional vector of random coefficients and $\boldsymbol{\epsilon}_i$ is the vector of errors. Fixed effects are identified by parameters that are associated with the entire population, whereas random effects are identified by group-specific parameters.

In the parametric framework of mixed effects linear models, random coefficients are assumed to be distributed according to a normal distribution with unknown parameters that, together

with the coefficients of fixed effects and σ^2, can be estimated through methods based on the maximization of the likelihood or the restricted likelihood functions (Pinheiro and Bates, 2000).

The main novelty that is introduced here is that we move to a semiparametric framework, assuming that the coefficients $\mathbf{b}_i$ are distributed according to a discrete distribution P^*, assuming M sets of values $(c_{0l}, \ldots, c_{rl})$ for $l = 1, \ldots, M$, where $M \leqslant N$. This means that each group i, for $i = 1, \ldots, N$, is assigned to a subpopulation l that is characterized by random parameters $(c_{0l}, \ldots, c_{rl})$. This semiparametric modelling enables us to identify a latent structure among the groups, that are clustered by the model into an unknown number of discrete masses. Therefore, the two main advantages are that, firstly, we can identify how many latent subpopulations exist within the groups of data and, second, we can estimate the parameters that are associated with each subpopulation, pointing out their differences.

Under these assumptions, the semiparametric mixed effects model takes the form

$$\begin{gathered} \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{c}_l + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N, \quad l = 1, \ldots, M, \\ \boldsymbol{\epsilon}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{1}_{n_i}). \end{gathered} \tag{2}$$

In particular, from now on, without loss of generality, we consider the case with one random intercept, one random effect and one fixed effect:

$$\begin{gathered} \mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{1}c_{0l} + \mathbf{z}_i c_{1l} + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N, \quad l = 1, \ldots, M, \\ \boldsymbol{\epsilon}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{1}_{n_i}) \end{gathered} \tag{3}$$

where $\mathbf{1}$ is the n_i-dimensional vector of 1s and $M \leqslant N$ is the number of subpopulations (mass points) which are unknown *a priori*. (This choice of model is due to the case that is considered in the application to the INVALSI data set, in Section 3.) Coefficients $\mathbf{c}_l$, for $l = 1, \ldots, M$, are distributed according to a probability measure $\mathcal{P}^*$ that belongs to the class of all probability measures on $\mathbb{R}^2$. $\mathcal{P}^*$ is a discrete measure with M support points that can then be interpreted as the mixing distribution that generates the density of the stochastic model (3). The maximum likelihood estimator $\hat{\mathcal{P}}^*$ of $\mathcal{P}^*$ can be obtained following the theory of mixture likelihoods in Lindsay (1983a, b), 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 (1983a, b) 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. So, the maximum likelihood estimator of the random-effects distribution can be expressed as a set of points $(\mathbf{c}_1, \ldots, \mathbf{c}_M)$, where $M \leqslant N$ and $\mathbf{c}_l \in \mathbb{R}^2$ for $l = 1, \ldots, M$, and a set of weights $(w_1, \ldots, w_M)$, where $\Sigma_{l=1}^{M} w_l = 1$ and $w_l \geqslant 0$ for each $l = 1, \ldots, M$. Given this, we propose an algorithm for the joint estimation of σ^2, β, $(\mathbf{c}_1, \ldots, \mathbf{c}_M)$ and $(w_1, \ldots, w_M)$, which is performed through the maximization of the likelihood, by the discrete distribution of the random effects,

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = \sum_{l=1}^{M} \frac{w_l}{(2\pi\sigma^2)^{J/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} (y_{ij} - \beta x_{ij} - c_{0l} - c_{1l} z_{ij})^2 \right\}, \tag{4}$$

with respect to the fixed coefficient β, the error variance σ^2 and the random-effects distribution $(\mathbf{c}_l, w_l)$, for $l = 1, \ldots, M$. For each $l = 1, \ldots, M$, $\mathbf{c}_l$ represents the group-specific parameters and w_l the corresponding weight in the mixture equation (3).

The algorithm that we propose is inspired by that proposed in Azzimonti *et al.* (2013), but it considers the linear functional dependence between response and predictors and it makes three main improvements:

(a) the optimization of the maximization step is computed in closed form,
(b) the covariates can be group specific and
(c) the initialization of the parameters is done in a more efficient and flexible way

(by the term 'group-specific covariates' we mean individual level covariates that are allowed to vary in terms of the number of observations and assumed values across the groups). The first point directly derives from the linearity assumption. The idea at the base of the algorithm is also similar to that proposed in Aitkin (1996) but, whereas Aitkin (1996) needed to fix *a priori* the number of discrete points of the mixing distribution, our algorithm itself identifies the number of support points M, on the basis of given tolerance values that we fix depending on the problem.

2.2. The semiparametric expectation–maximization algorithm

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 answer variable y_{ij} and of the covariates z_{ij} and x_{ij}, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$. The parameters to be estimated are the random coefficients $\mathbf{c}_l$ with their weights w_l, for $l = 1, \ldots, M$, the fixed coefficient β and the variance σ^2. The algorithm allows the number n_i, for $i = 1, \ldots, N$, of observations to be different across groups, but, within each group missing data are not handled, i.e. missing values of y, z and x for the n_i units are not allowed. At each iteration, the EM algorithm updates the parameters to increase the likelihood in equation (4) and it continues until convergence or until a fixed number of iterations, it, is reached. In particular, the update is given by

$$w_l^{(\mathrm{up})} = \frac{1}{N} \sum_{i=1}^{N} W_{il} \qquad l = 1, \ldots, M, \tag{5}$$

$$(\boldsymbol{\beta}^{(\mathrm{up})}, \mathbf{c}_1^{(\mathrm{up})}, \ldots, \mathbf{c}_M^{(\mathrm{up})}, \sigma^{2(\mathrm{up})}) = \arg\max_{\boldsymbol{\beta}, \mathbf{c}_l, \sigma^2} \sum_{l=1}^{M} \sum_{i=1}^{N} W_{il} \ln\{p(\mathbf{y}_i | \boldsymbol{\beta}, \sigma^2, \mathbf{c}_l)\} \tag{6}$$

where

$$W_{il} = \frac{w_l p(\mathbf{y}_i | \boldsymbol{\beta}, \sigma^2, \mathbf{c}_l)}{\sum_{k=1}^{M} w_k p(\mathbf{y}_i | \boldsymbol{\beta}, \sigma^2, \mathbf{c}_k)} \tag{7}$$

and

$$p(\mathbf{y}_i | \boldsymbol{\beta}, \sigma^2, \mathbf{c}) = \frac{1}{(2\pi\sigma^2)^{n_i/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} (y_{ij} - \beta x_{ij} - c_{0l} - c_{1l} z_{ij})^2 \right\}. \tag{8}$$

The weight $w_l^{(\mathrm{up})}$ is the mean over the N groups of their weights related to the lth subpopulation. Coefficients W_{il} represent the probability that $\mathbf{b}_i$ is equal to $\mathbf{c}_l$ conditionally on observations $\mathbf{y}_i$ and given the fixed coefficient β and the variance σ^2.

The maximization step in equation (6) involves two steps and it is done iteratively. In the first step, we compute arg max with respect to the support points $\mathbf{c}_l$, keeping β and σ^2 fixed to the last computed values. In this way, we can maximize the expected log-likelihood (computed in the expectation step) with respect to all support points $\mathbf{c}_l$ separately, i.e.

$$\mathbf{c}_l^{(\mathrm{up})} = \arg\max_{\mathbf{c}} \sum_{i=1}^{N} W_{il} \ln\{p(\mathbf{y}_i|\boldsymbol{\beta}, \sigma^2, \mathbf{c})\} \qquad l = 1, \ldots, M. \tag{9}$$

Since we are considering the linear case, it is possible to perform this maximization step in closed form. With regard to model (3), the estimates of the random effects are obtained by means of the weighted least squares method as follows:

$$\hat{c}_{0l} = \frac{\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta} x_{ij} - \hat{c}_{1l} z_{ij})}{n_i \sum_{i=1}^{N} w_{il}} \tag{10}$$

and

$$\begin{aligned}\hat{c}_{1l} = {} & \frac{\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} y_{ij} z_{ij} - \frac{\left(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} y_{ij}\right)\left(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij}\right)}{n_i \sum_{i=1}^{N} w_{il}}}{\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij}^2 - \frac{\left(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij}\right)^2}{n_i \sum_{i=1}^{N} w_{il}}} \\ & + \frac{\frac{\hat{\beta}\left(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij}\right)\left(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} x_{ij}\right)}{n_i \sum_{i=1}^{N} w_{il}} - \hat{\beta} \sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} x_{ij} z_{ij}}{\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij}^2 - \frac{\left(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij}\right)^2}{n_i \sum_{i=1}^{N} w_{il}}}.\end{aligned} \tag{11}$$

In the second step, we fix the support points of the random-effects distribution computed in the previous step and we compute arg max in equation (6) with respect to β and σ^2. Again, this step can be done in closed form and the estimates of the parameters, with regard to model (3), obtained by means of the weighted least squares method, are

$$\hat{\beta} = \frac{\sum_{l=1}^{M} \sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} (y_{ij} x_{ij} - \hat{c}_{0l} x_{ij} - \hat{c}_{1l} z_{ij} x_{ij})}{\sum_{l=1}^{M} \sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} x_{ij}^2} \tag{12}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{l=1}^{M} \sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta} x_{ij} - \hat{c}_{0l} - \hat{c}_{1l} z_{ij})^2}{n_i \sum_{l=1}^{M} \sum_{i=1}^{N} w_{il}}. \tag{13}$$

Since $w_l = p(\mathbf{b}_i = \mathbf{c}_l)$, then

$$W_{il} = \frac{w_l\, p(\mathbf{y}_i|\boldsymbol{\beta}, \sigma^2, \mathbf{c}_l)}{\sum_{k=1}^{M} w_k\, p(\mathbf{y}_i|\boldsymbol{\beta}, \sigma^2, \mathbf{c}_k)} = \frac{p(\mathbf{b}_i = \mathbf{c}_l)\, p(\mathbf{y}_i|\beta, \sigma^2, \mathbf{c}_l)}{p(\mathbf{y}_i|\beta, \sigma^2)}$$
$$= \frac{p(\mathbf{y}_i, \mathbf{b}_i = \mathbf{c}_l|\beta, \sigma^2)}{p(\mathbf{y}_i|\beta, \sigma^2)}$$
$$= p(\mathbf{b}_i = \mathbf{c}_l|\mathbf{y}_i, \beta, \sigma^2). \tag{14}$$

Therefore, to compute the point $\mathbf{c}_l$ for each group i, for $i = 1, \ldots, N$, we maximize the conditional probability of $\mathbf{b}_i$ given the observations $\mathbf{y}_i$, the coefficient β and the error variance σ^2. The estimate of the coefficients $\mathbf{b}_i$ of the random effects for each group is obtained by maximizing W_{il} over l, i.e.

$$\hat{\mathbf{b}}_i = \mathbf{c}_{\tilde{l}} \qquad \tilde{l} = \arg\max_{l} W_{il} \quad i = 1, \ldots, N. \tag{15}$$

As anticipated before, the initialization of the support points is done in a robust and generalizable way. The algorithm starts by considering N support points for the coefficients of random effects and a starting estimate for the coefficients of fixed effects. In particular, the initialization of all these parameters is done in the following way.

(a) Random effects: the starting N support points are obtained by fitting a simple linear regression within each group and estimating the pair of parameters (both the intercept and the slope) for each of the N groups. The weights are uniformly distributed on these N support points. (This is not the only possibility to estimate the starting support points. A valuable alternative is to fit a classical multilevel model, with N groups, where both the intercept and the slope are random coefficients.)
(b) Fixed effects: the starting values of β and σ^2 are estimated by fitting a unique linear regression on the entire population (without distinction among the groups).

Nonetheless, if the number of starting support points N is extremely large, the algorithm is relatively slow and using N starting support points becomes not strictly necessary. In this case, the initialization of the support points of the random-effect distribution is done in the following way:

(a) we choose a number $N^* < N$ of support points;
(b) we extract N^* points from a uniform distribution with support on the entire range of possible values, i.e. estimated by fitting N distinct linear regressions for each of the N groups, as before, and identifying the minimum and the maximum values;
(c) we uniformly distribute the weights on these N^* support points.

During the iterations, the EM algorithm performs the support reduction of the discrete distribution, to identify $M < N$ mass points in which the N groups are clustered. The support reduction is made on the basis of two criteria. The first is that we fix a threshold D and if two points $\mathbf{c}_l$ and $\mathbf{c}_k$ are closer than D, in terms of Euclidean distance, they collapse to a unique point $\mathbf{c}_{l,k}$, where $\mathbf{c}_{l,k} = (\mathbf{c}_l + \mathbf{c}_k)/2$ with weight $w_{l,k} = w_l + w_k$. 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, and so on. The second is that, starting from a given iteration up to the end, we fix a threshold $\tilde{w}$ and we remove mass points with weight $w_l \leqslant \tilde{w}$ or that are not associated with any subpopulation. D and $\tilde{w}$ are two tuning parameters that tune the estimates

of the subpopulations. The choice of D depends on how much we want to be sensitive to the differences between subpopulations: the higher is D, the lower is the number of subpopulations and the less homogeneous are the groups within subpopulations. D depends also on the order of magnitude of the data. The choice of $\tilde{w}$ depends on the minimum number of groups that we allow within 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:

$$\begin{aligned} S_w &= \sum_{l=1}^{M^{\text{new}}} w_l^{\text{old}}, \\ w_l^{\text{new}} &= \frac{w_l^{\text{old}}}{S_w} \qquad \forall\, l = 1, \ldots, M^{\text{new}} \end{aligned} \tag{16}$$

where M^{new} is the total number of masses after deleting those associated with weight $w_l \leqslant \tilde{w}$ or not associated with any subpopulation, $\mathbf{w}^{\text{old}}$ are the old remaining weights and $\mathbf{w}^{\text{new}}$ are the new reparameterized weights.

A sketch of the algorithm is shown in algorithm 1 in Table 1. At each iteration k, the algorithm, given the estimated number of mass points, estimates all the parameters in equation (3) in an iterative way, updating the coefficients of both fixed and random effects, until convergence or until it reaches 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 $\tilde{w}$. 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.

The introduction of the maximum number of iterations it, it1 and itmax (as just explained) depends on the complexity of the data and on the consequent rate of convergence and its use is merely to avoid an infinite loop.

It is worth noting that, since the optimization steps are done in closed form, the algorithm is not particularly time consuming and, in both the simulation study and in the application, it converged in fewer than 20 iterations.

In the presentation of the algorithm, as well as in the simulation study that will be presented in the next subsection, we focus on the case of a linear model with two covariates, where both one slope and the intercept are considered as random effects. This is due to the upcoming application of the algorithm to the case-study of the INVALSI data set, in which we make this choice of fixed and random parameters. Nonetheless, the semiparametric EM algorithm enables us to consider as random effects both the intercept and one slope, as well as only one of them. Moreover, its extension to the case with p covariates among the random effects, i.e. $\mathbf{c} \in \mathbb{R}^{p+1}$, is analytically straightforward and it implies only a computational issue.

2.3. Simulation study

To validate the estimation algorithm proposed, we perform two simulation studies: the first considers the case of a population containing three latent subpopulations and the second considers the case of a population with no latent subpopulations. In this way, we can test the algorithm

Table 1. Algorithm 1—EM algorithm for semiparametric mixed effects models

Input: initial estimates for $(\mathbf{c}_1^{(0)}, \ldots, \mathbf{c}_M^{(0)})$ and $(w_1^{(0)}, \ldots, w_M^{(0)})$, with $M = N$; initial estimates for $\beta^{(0)}$ and $\sigma^{2(0)}$; tolerance parameters D, $\tilde{w}$, tollR, tollF, it, it1 and itmax
Output: final estimates of $\mathbf{c}_l^{(it)}$, $w_l^{(it)}$, for $l = 1, \ldots, M$, $\beta^{(it)}$ and $\sigma^{2(it)}$
$k = 1$; conv1 $= 0$; conv2 $= 0$
While (conv1 $==0$ or conv2 $==0$ & $k < it$) *do*
  compute the distance matrix DIST (where $\text{DIST}_{st} = \sqrt{\{(c_{0s} - c_{0t})^2 + (c_{1s} - c_{1t})\}}^2$ is the Euclidean distance between each couple of mass points $s, t (\forall\, s, t = 1, \ldots, M, s \neq t)$;
  if ($\text{DIST}_{st} < D$ & $\text{DIST}_{st} = \min(\text{DIST})$ $(\forall\, s, t = 1, \ldots, M, s \neq t)$) *then*
    collapse masses s and t to a unique mass point;
    compute the new distance matrix DIST;
  if conv1 $==1$ or $k \geqslant$ it1 *then*
    if $w_l^{(k)} \leqslant \tilde{w}$ $(\forall\, l = 1, \ldots, M)$ *then*
      delete mass point l;
      reparameterize the weights according to equation (16);
    if no changes are done *then*
      conv2 $= 1$;
  given $\mathbf{c}_l^{(k-1)}, w_l^{(k-1)}$ for $l = 1, \ldots, M$, $\beta^{(k-1)}$ and $\sigma^{2(k-1)}$, compute the matrix W according to equation (7);
  update the weights $w_1^{(k)}, \ldots, w_M^{(k)}$ according to equation (5);
  $\beta^{(k,0)} = \beta^{(k-1)}$;
  $\sigma^{2(k,0)} = \sigma^{2(k-1)}$;
  $\mathbf{c}_l^{(k,0)} = \mathbf{c}_l^{(k-1)}$;
  $w_l^{(k,0)} = w_l^{(k-1)}$;
  keeping $\beta^{(k,0)}$ and $\sigma^{2(k,0)}$ fixed, update the M support points $\mathbf{c}_1^{(k,1)}, \ldots, \mathbf{c}_M^{(k,1)}$ according to equations (10) and (11);
  keeping $\mathbf{c}_l^{(k,1)}, w_l^{(k,0)}$ for $l = 1, \ldots, M$ fixed, update $\beta^{(k,1)}$ and $\sigma^{2(k,1)}$ according to equations (12) and (13);
  $j = 1$;
  while ($|\beta^{(k,j-1)} - \beta^{(k,j)}| \geqslant$ tollF or $|\sigma^{2(k,j-1)} - \sigma^{2(k,j)}| \geqslant$ tollF or $|\mathbf{c}_l^{(k,j-1)} - \mathbf{c}_l^{(k,j)}| \geqslant$ tollR) & $j \leqslant$ itmax *do*
    $j = j + 1$;
    keeping $\beta^{(k,j-1)}$ and $\sigma^{2(k,j-1)}$ fixed, update the M support points $\mathbf{c}_1^{(k,j)}, \ldots, \mathbf{c}_M^{(k,j)}$ according to equations (10) and (11);
    keeping $\mathbf{c}_l^{(k,j)}, w_l^{(k,j-1)}$ for $l = 1, \ldots, M$ fixed, update $\beta^{(k,j)}$ and $\sigma^{2(k,j)}$ according to equations (12) and (13);
  set $\mathbf{c}_l^{(k)} = \mathbf{c}_l^{(k,j)}$ for $l = 1, \ldots, M$, $\beta^{(k)} = \beta^{(k,j)}$, $\sigma^{2(k)} = \sigma^{2(k,j)}$;
  estimate subpopulation l for each group i according to equation (15);
  if $(\beta^{(k)} - \beta^{(k-1)} <$ tollF) & $(\sigma^{2(k)} - \sigma^{2(k-1)} <$ tollF) & $(\mathbf{c}_l^{(k)} - \mathbf{c}_l^{(k-1)} <$ tollR) *then*
    conv1 $= 1$;
  $k = k + 1$

in the presence of distinct subpopulations and also in the case in which there are no distinct subpopulations. We apply the algorithm considering various values of D, to test how the results change by changing the threshold parameter and we provide a measure of the uncertainty of classification by computing the entropy in the weights matrix W. We consider a linear model with two covariates.

For the first simulation study, we generate a data set containing 100 groups of variables (100 level 2 units), where each group is composed of an answer variable and two covariates. We sample the variables to have three different latent subpopulations within the 100 groups, i.e. to create 100 cohorts of data characterized by three different linear correlations. For this, we generate 100 response variables as the result of three distinct linear combinations of three pairs of covariates, plus some errors. The three subpopulations contain 40, 25 and 35 groups. The data are simulated in the following way:

$$\left.\begin{aligned} \mathbf{y}_i &= \beta \mathbf{x}_1 + c_{01} + c_{11}\mathbf{z}_1 + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, 40, \\ \mathbf{y}_i &= \beta \mathbf{x}_2 + c_{02} + c_{12}\mathbf{z}_2 + \boldsymbol{\epsilon}_i \qquad i = 41, \ldots, 65, \\ \mathbf{y}_i &= \beta \mathbf{x}_3 + c_{03} + c_{13}\mathbf{z}_3 + \boldsymbol{\epsilon}_i \qquad i = 66, \ldots, 100 \end{aligned}\right\} \tag{17}$$

where the coefficients β and $\mathbf{c}_l$, for $l = 1, \ldots, 3$, are reported in Table 2, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, 3)$ and the covariates are sampled from normal distributions with different parameters. In particular,

$$\left.\begin{aligned} \mathbf{x}_1 &\sim \mathcal{N}(0.30, 0.16), \qquad \mathbf{z}_1 \sim \mathcal{N}(50, 100), \\ \mathbf{x}_2 &\sim \mathcal{N}(0.28, 0.16), \qquad \mathbf{z}_2 \sim \mathcal{N}(51, 100), \\ \mathbf{x}_3 &\sim \mathcal{N}(0.27, 0.16), \qquad \mathbf{z}_3 \sim \mathcal{N}(49, 100), \end{aligned}\right\} \tag{18}$$

where $\mathbf{z}_1$ and $\mathbf{x}_1$ have 100 observations, $\mathbf{z}_2$ and $\mathbf{x}_2$ have 90 observations and $\mathbf{z}_3$ and $\mathbf{x}_3$ have 95 observations (9575 level 1 units in total). Therefore, the dimensional choices of the data generated are as follows:

$$\text{number of groups} = 100,$$

$$\text{number of subjects within groups} = \begin{cases} 100 & \forall \text{group } i \in \{1, \ldots, 40\}, \\ 90 & \forall \text{group } i \in \{41, \ldots, 65\}, \\ 95 & \forall \text{group } i \in \{66, \ldots, 100\}. \end{cases}$$

The choice of the size, of the parameters and of the distribution is arbitrary. Our choice for the values of x and z is driven by the case-study. We sample x and z to obtain values in the same range as those in the INVALSI application. Other choices are possible and do not affect the validity of the results.

Also the choice of the coefficients in Table 2 is arbitrary. This choice of parameters is driven by the case-study, since we choose values for $\mathbf{c}_l$, for $l = 1, \ldots, 3$ and β in the same range as those obtained in the INVALSI application. For coherence with the upcoming INVALSI case-study, which considers both the slope and the intercept as random, we choose different values for both the intercept and the coefficient of variable $\mathbf{z}$ across the three subpopulations, whereas we hold the coefficient of $\mathbf{x}$ fixed. Fig. 1 shows the three-dimensional image of one simulated data set.

Looking at Fig. 1, it is possible to recognize three different linear correlations between the data, identified by the three distinct 'clouds' of points. Groups of points that are characterized by similar linear correlations are automatically associated with similar colours by the R software and this helps in the visual inspection of the three subpopulations.

Table 2. Coefficients used for data simulation in equation (17)†

l	c_0	c_1	β
1	20	1.00	1.50
2	30	0.05	1.50
3	40	0.50	1.50

†Each row corresponds to a subpopulation l. The intercept and the coefficient of $\mathbf{z}$ differ across subpopulations (c_0 and c_1 respectively), whereas the coefficient of x (β) is fixed.

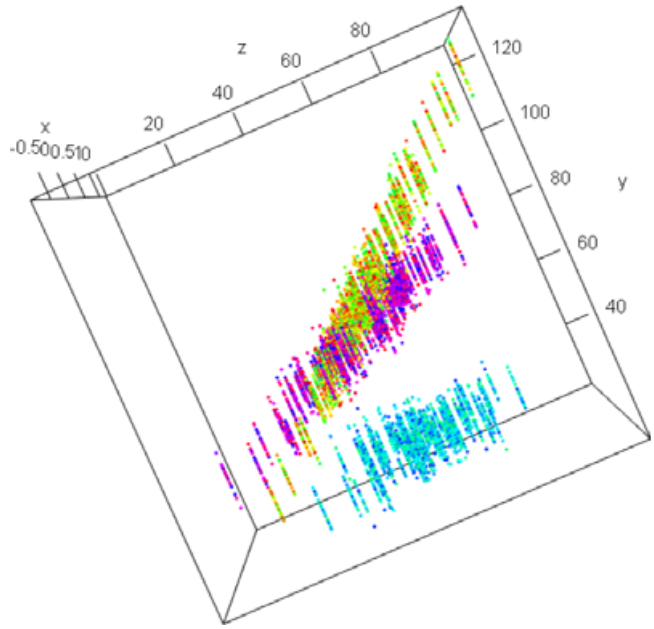


Fig. 1. Plot of the simulated data obtained by equations (17) and (18): each of the 100 groups has a different colour; data with similar behaviours are automatically assigned to similar colours by the R software

The model that we fit takes the form

$$\mathbf{y}_i = \beta \mathbf{x}_i + c_{0l} + c_{1l}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \tag{19}$$

where $i = 1, \ldots, 100$ and $l = 1, \ldots, M$ where M is unknown *a priori* in the algorithm. We apply the algorithm 100 times, to different simulated data sets for the same model, for each different value of $D = \{0.5, 0.8, 1, 2, 3\}$ and considering the following choice of the other parameters: $\tilde{w} = 0.05$, it $= 30$, it1 $= 20$, itmax $= 20$ and toolF $=$ tollR $= 10^{-4}$. The following scheme summarizes the simulation study:

$$\forall\, D \in \{0.5, 0.8, 1, 2, 3\} \text{ and for } (k \text{ in } 1:100),$$

generate $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$ according to equation (18) and $\mathbf{y}$ according to equation (17); apply the semiparametric EM algorithm to the generated data.

The number of times, out of the 100 runs, in which the algorithm allocates the right subpopulation to each of the 100 groups, for various values of D, is shown in Table 3 (in all the runs, the algorithm converges before the maximum number of iterations).

In the case in which $D = 0.5$, the algorithm correctly assigns groups to the three subpopula-

Table 3. Number of times, out of the 100 runs, in which the algorithm allocates the right subpopulation to each of the 100 groups for various values of D

$D=0.5$	$D=0.8$	$D=1$	$D=2$	$D=3$
34	84	92	98	68

tions 34 times out of 100. In the remaining 66 cases, the algorithm identifies more than three subpopulations. This means that the threshold value $D = 0.5$ is too small and the algorithm is, consequently, too sensitive to the variations in the data. In contrast, in the case in which $D = 3$, the algorithm correctly assigns the groups to the three subpopulations 68 times out of 100 (identifying fewer than three subpopulations in the remaining 32 cases). This means that, for values of D higher than 3, the algorithm is not perfectly sensitive to the differences between the groups and it sometimes collapses groups presenting different trends into the same subpopulation. In the cases of $D = \{0.8, 1, 2\}$ the algorithm correctly assigns the subpopulations 84, 92 and 96 times out of 100 respectively, which represents a good proportion. The results of the estimates of the parameters for the two 'best' choices of D are shown in Table 4.

Starting from 100 distinct groups, the semiparametric EM algorithm, in most of the cases, identifies three subpopulations ($M = 3$) that are represented by the estimates $(\hat{\mathbf{c}}_l, \hat{w}_l)$, for each $l = 1, \ldots, M$, and $\hat{\beta}$ shown in Table 4. The estimates that are obtained with $D = 1$ and $D = 2$ are coherent. The mean of each parameter distribution is centred very close to the real value of the parameter that was used to simulate the data and the standard deviations are very small. (To test equality of the mean of each parameter distribution to the parameters shown in Table 2, we test the normality of each parameter distribution by means of a Shapiro test, obtaining p-values that are greater than 0.1 for all of them, and we perform a t-test for each parameter (c_0, c_1 and β), obtaining p-values that are greater than 0.2 for all the tests.) Moreover, the masses' volumes are proportional to the percentage of data that belongs to each mass. In this case, the algorithm correctly assigns the 100 groups to the three subpopulations, so that the three volumes are proportional to 0.40, 0.25 and 0.35. For one of the 100 simulated data sets in which the algorithm identifies the three clusters, data with the three regression planes identified are shown in Fig. 2.

In this simulation, the algorithm associates each observation with the correct subpopulation. The three regression planes identified can fit the three distinct clouds of data in a precise way. To have a measure of the uncertainty of classification of the semiparametric EM algorithm,

Table 4. Distribution of the parameters of model (19), estimated by the semiparametric EM algorithm, obtained in the runs in which three populations are identified†

l	$\hat{c}_0$		$\hat{c}_1$		$\hat{\beta}$		$\hat{w}$
	Mean	*Standard deviation*	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*	
$D = 1$							
1	20.034	0.170	0.999	0.003			0.40
2	40.001	0.197	0.500	0.003	1.477	0.005	0.25
3	30.032	0.292	0.049	0.005			0.35
$D = 2$							
1	20.011	0.154	1.000	0.003			0.40
2	40.038	0.176	0.499	0.004	1.505	0.004	0.25
3	29.987	0.236	0.050	0.004			0.35

†Results are shown both for $D = 1$ and for $D = 2$. Within each choice of D, each row corresponds to a subpopulation l. The intercept and the coefficient of $\mathbf{z}$ differ across subpopulations (c_0 and c_1 respectively), whereas the coefficient of $\mathbf{x}$ (β) is fixed. $\hat{w}$ represents the weight estimated for each subpopulation.

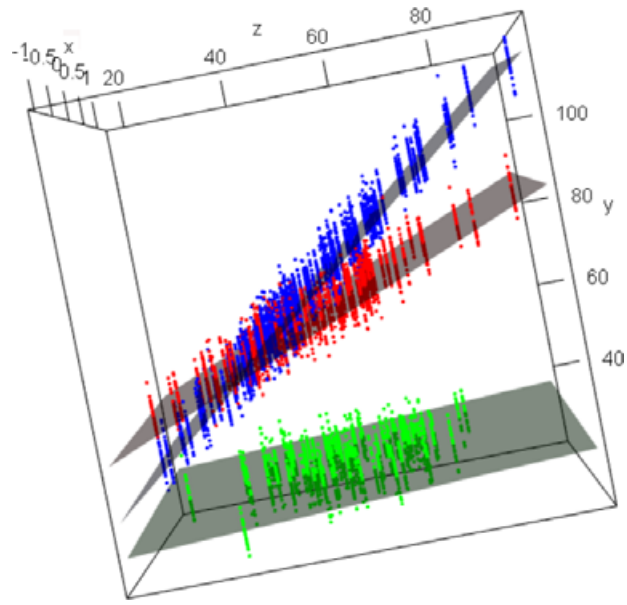


Fig. 2. Result of the semiparametric EM algorithm applied to a simulated data set according to equations (17) and (18): colours represent the three subpopulations that the algorithm identifies and planes are the estimated linear regression planes within each subpopulation; each group is painted with the colour of the subpopulation to which it belongs

Table 5. Mean and standard deviation of the entropy estimated when $D = 0.8$, $D = 1$ and $D = 2$ on the 100 runs of the simulation, for the choice of data in equation (18) and coefficients in Table 2

D	*Mean*	*Standard deviation*
0.8	0.019	0.012
1	0.013	0.008
2	9.7×10^{-13}	9.6×10^{-12}

we can observe the matrices of weights W that we obtain in each run and evaluate the level of uncertainty with which the algorithm assigns each group to a cluster. This uncertainty of classification can be evaluated by measuring the entropy of the rows of the matrix W. In the best case, i.e. when the algorithm assigns each group i to a cluster l with probability 1, each row of the matrix W would be composed of $M-1$ values equal to 0 and a value equal to 1. In this scenario, the entropy $E_i = -\Sigma_{l=1}^{M} W_{il} \ln(W_{il})$ of each row i of the matrix W would be equal to 0. The more the distribution of the weights is uniform on the M mass points, the higher is the entropy. The worst case when $M = 3$ is the case in which the distribution of the weights of a group i is uniform on the three clusters ($w_{il} = \frac{1}{3}$ for $l = 1, 2, 3$), which corresponds to an entropy $E_i = -3 \times \frac{1}{3} \ln(\frac{1}{3}) = 1.098$. We compute the entropy of each row of W for the 100 runs and we show here the distribution of the mean on the 100 runs of the entropy measured for each group i, in the cases of $D = 0.8$, $D = 1$ and $D = 2$.

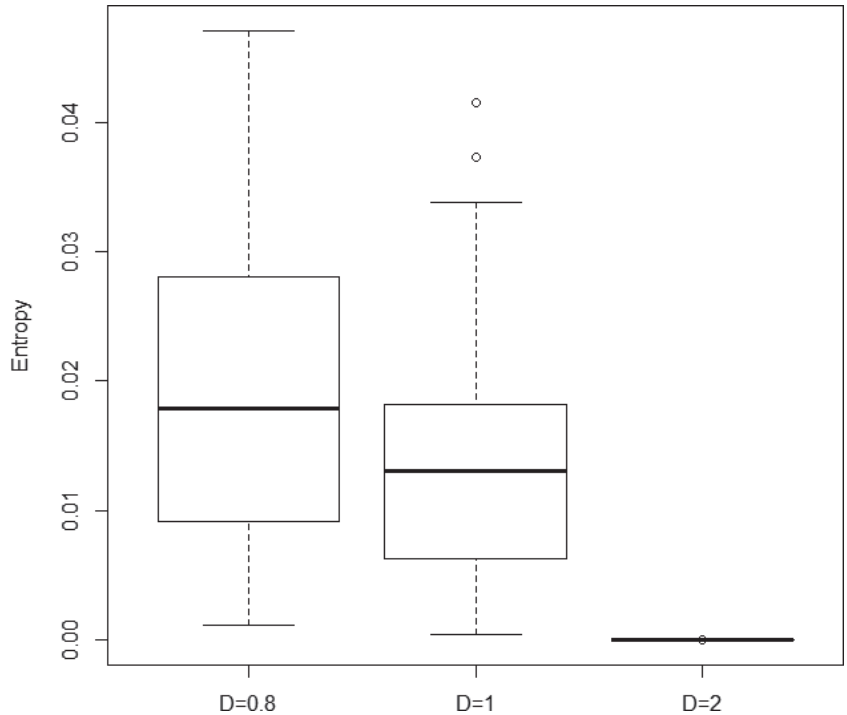


Fig. 3. Boxplots of the entropy computed in the 100 runs, for $D = 0.8$, $D = 1$ and $D = 2$: each boxplot represents the distribution of the entropy measured for each group, obtained by mediating the entropy in the 100 runs

Table 6. Number of times, out of the 100 runs, in which the algorithm identifies only one subpopulation, for various values of D

$D=0.5$	$D=0.8$	$D=1$	$D=2$	$D=3$
52	74	90	100	100

The mean and the standard deviation of the entropy that are estimated when $D=0.8$, $D=1$ and $D=2$ are shown in Table 5.

These very low values of the entropy (Fig. 3 and Table 5) suggest that the level of uncertainty of classification, for these three values of D, is very low, since the distribution of the weights w_{il}, for $i=1,\ldots,N$ and $l=1,\ldots,3$, turns out to be very concentrated on single mass points. In particular, the case in which $D=2$ has the lowest entropy and turns out to be the case with the lowest level of uncertainty of classification.

We can conclude that, in this simulation study, the semiparametric EM algorithm can identify the latent structure that elapses within the 100 groups of data. In particular, it can identify which is the effective number of subpopulations in which the data are nested and it can characterize each of these subpopulations by means of the estimates of the associated parameters.

In the second simulation study, we generate a population without latent subpopulations and we analyse the performance of the semiparametric EM algorithm. We choose one of the previous

Table 7. Distribution of the parameters of model (20), estimated by the semiparametric EM algorithm, obtained in the 100 runs†

l	$\hat{c}_0$		$\hat{c}_1$		$\hat{\beta}$		$\hat{w}$
	Mean	*Standard deviation*	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*	
1	20.012	0.099	0.999	0.002	1.493	0.081	1

†Results are shown for $D = 2$ but are coherent with any other choice of $D \geqslant 2$.

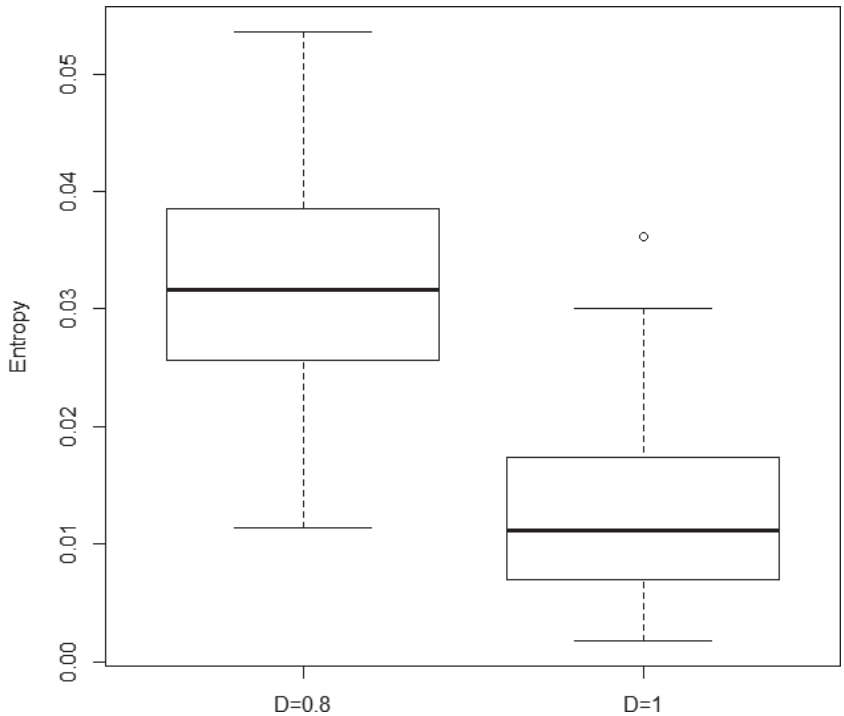


Fig. 4. Boxplots of the entropy computed in the 100 runs, for $D = 0.8$ and $D = 1$: each boxplot represents the distribution of the entropy measured for each group, obtained by mediating the entropy in the 100 runs

set of parameters and we generate 100 response variables in the following way:

$$\mathbf{y}_i = 20 + 1.5\mathbf{x}_i + 1\mathbf{z}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, 100, \tag{20}$$

where $\epsilon_i \sim \mathcal{N}(0, 3)$ and x_i and z_i are defined as in equation (18). Again, we apply the algorithm 100 times to 100 different simulated data sets and this process is repeated for values of $D = \{0.5, 0.8, 1, 2, 3\}$ and considering the following choice of the other parameters: $\tilde{w} = 0.05$, it $= 30$, it1 $= 20$, itmax $= 20$ and toolF $=$ tollR $= 10^{-4}$. The number of times, out of the 100 runs, in which the algorithm identifies only one subpopulation, for various values of D, is shown in Table 6.

For $D = 2$ and $D = 3$, the algorithm always recognizes that there are no subpopulations. For smaller values of D, sometimes the algorithm catches heterogeneities among the 100 groups

Table 8. Mean and standard deviation of the entropy estimated when $D = 0.8$ and $D = 1$ on the 100 runs of the simulation, for the choice of coefficients in equation (20)

D	*Mean*	*Standard deviation*
0.8	0.032	0.008
1	0.012	0.007

of data and identifies the presence of latent subpopulations. For $D = 2$, Table 7 shows the distribution of the estimated coefficients in the 100 runs.

Regarding the uncertainty of classification, the entropy in the simulations with $D = 2$ and $D = 3$ is 0, since 100 times out of 100 the algorithm identifies one population and each group has probability 1 assigned to it. In the cases of lower values of D, the algorithm sometimes identifies more than one population and the distribution of the entropy related to these cases is shown in Fig. 4.

The mean and the standard deviation of the entropy that are estimated when $D = 0.8$ and $D = 1$ are shown in Table 8.

Also in this case, the estimates of the parameters turn out to be significantly equal to the parameters that were used to generate the data. (Again, we test the normality of each parameter distribution, obtaining p-values of the Shapiro test greater than 0.1 for all of them, and t-tests for the null hypotheses $c_0 = 5$, $c_1 = 3$ and $\beta = 10$ give p-values that are greater than 0.2.)

In general, by changing the value of D, we make the algorithm more or less sensitive to the heterogeneity among the groups of data, i.e. given both by the clustering induced by construction and by the remaining randomness in the model (e.g. by the error term). From this perspective, a graphical visualization of the results can help in the choice of D.

3. Case-study: application of the semiparametric expectation–maximization algorithm to Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione education data

In this section, we describe the INVALSI data set (Section 3.1) and we apply the semiparametric EM algorithm to these data, to identify subpopulations of Italian schools (Section 3.2). In the second step, we characterize the identified subpopulations by means of school level variables (Section 3.3).

3.1. The Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione data set

The INVALSI is an institute that tests Italian students at different grades and at different years. The data that we analyse in this paper are taken from the INVALSI survey of 2013–2014. Among others, the survey provides data both at student and at school level. Students take tests in different school subjects and fill out a questionnaire about themselves, their family situations and their habits. Moreover, school principals fill out a questionnaire about themselves, their school practices and management, body composition and school size, school structures, infra-

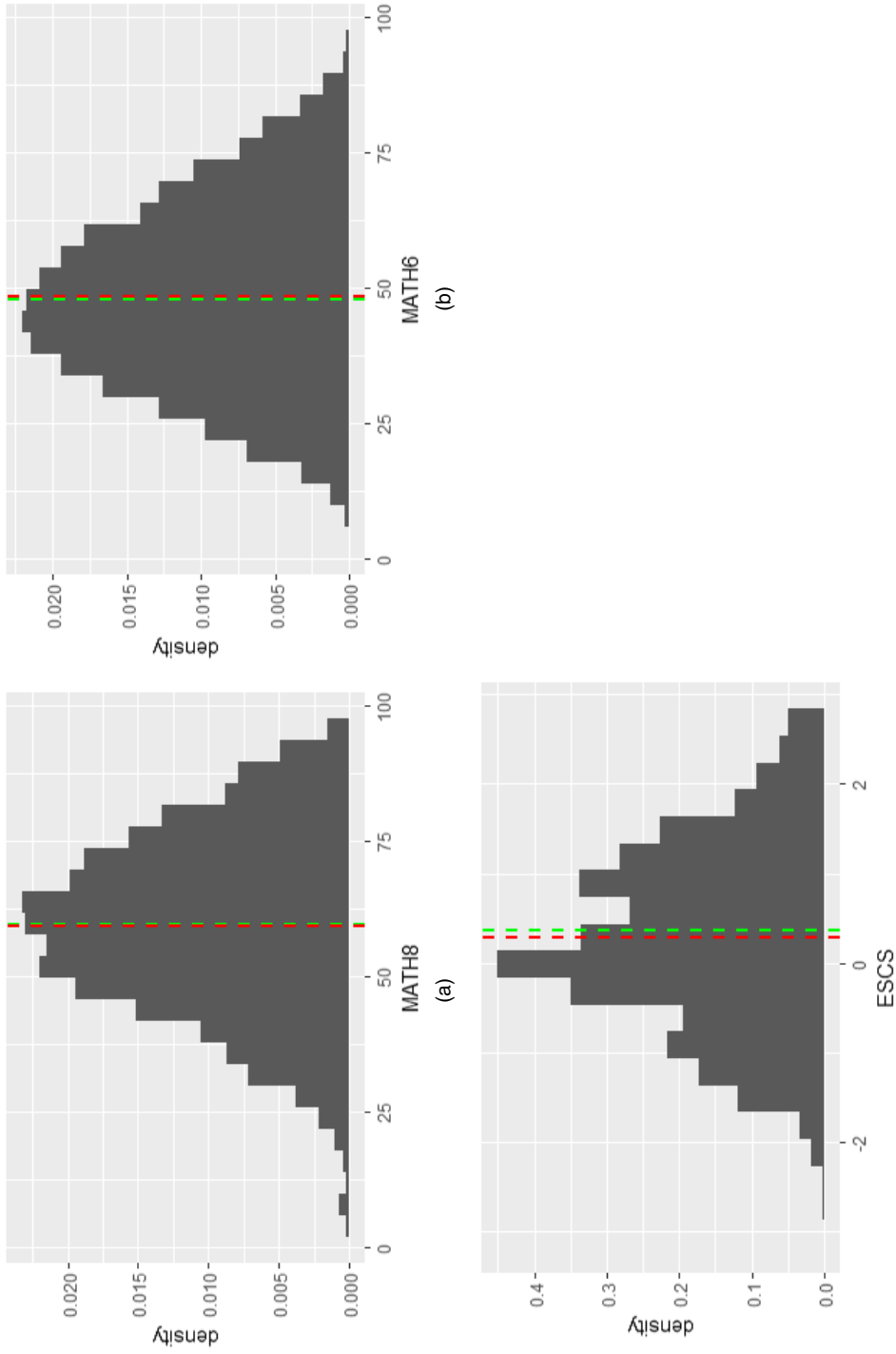


Fig. 5. Histograms of students' INVALSI test scores at (a) grade 8 and (b) grade 6 and (c) socio-economic index ESCS ⋮, means; ⋮, medians

Table 9. Descriptive statistics of student level variables employed in the analysis

	Mean	*Standard deviation*	*Median*	*Interquartile range*
MATH8	59.73	16.49	60.98	23.29
MATH6	48.69	16.83	48.26	24.55
ESCS	0.30	1.02	0.38	1.40

structures and school environment. The data set collects information about 8946 students nested within 586 schools. The aim of applying the semiparametric EM algorithm to the INVALSI data is that we are interested in exploring the different relationships between student performances at grade 6 and 8, across Italian junior secondary schools, adjusting for the student socio-economical index. For this reason, we select only three variables at student level to employ in the analysis:

(a) MATH8, student mathematics test score at grade 8 (students attending the last year of junior secondary school in the year 2013–2014);
(b) MATH6: student mathematics test score at grade 6 (students attending the first year of junior secondary school in the year 2011–2012);
(c) ESCS: a student socio-economic index.

Student test scores range between 0 and 100, whereas ESCS is an indicator built by the INVALSI as a continuous variable with mean 0 and variance 1. This indicator considers

(a) parents' occupation and educational qualifications and
(b) whether the student has certain items at home (for instance, the number of books).

In general, pupils with ESCS $\geqslant 2$ are socially and culturally highly advantaged. Fig. 5 and Table 9 show the variables' distributions and descriptive statistics respectively.

Moreover, we have information about the macroarea of localization of schools. About 59% of schools are in northern Italy, 18% are in central Italy and 23% are in southern Italy. Geographical information is very relevant since many studies in Italy confirm that there are significant discrepancies between student and school performances across the three geographical macroareas (Agasisti *et al.*, 2017; Agasisti and Vittadini, 2012; Masci *et al.*, 2016, 2017).

Since, in the second stage of the analysis, we shall look for a characterization of the school subpopulations identified, Table 10 reports the school level variables that we are interested in, with their descriptive statistics. In particular, the variables concern three characteristics of schools. The first concerns the *school body composition*: school mean socio-economic index, percentage of females, immigrants, late or early enrolled students, school size and a dummy for private or public school. (Late or early enrolled students are those students who started the school grade later or earlier with respect to their peers.) The second is about the *school principal's characteristics*: gender, age, education and years of experience. Lastly, we have three composite indicators about

(a) school environment and human relations,
(b) managerial practices and principal's strategy and
(c) structures and resources of the school.

(The computation of these three *composite indicators* is shown in Masci (2018).)

Table 10. School level variables of the database used in the analysis, with their descriptive statistics

Variable name	*Mean*	*Standard deviation*	*Median*	*Interquartile range*
Mean ESCS	0.26	0.54	0.27	0.58
Female percentage	50.11	10.83	50.00	14.28
Immigrant percentage	10.52	11.15	8.01	16.66
Early enrolled student percentage	1.21	4.13	0.00	0.00
Late enrolled student percentage	8.52	8.02	6.66	13.04
Number of classes	20.15	3.77	21.00	5.01
Number of school complexes	5.37	2.81	6.01	5.00
Private	8.21%	—	—	—
Principal features				
Gender (female $\equiv 1$)	70.01%	—	—	—
Age (years)	55.13	7.49	56.00	11.00
Master after degree (yes $\equiv 1$)	22%	—	—	—
Scientific education (yes $\equiv 1$)	14.62%	—	—	—
Years of experience	9.23	7.79	7.00	10.00
Years of experience in the actual school	5.08	5.18	3.00	5.00
Experience in another district	25.37%	—	—	—
Experience with INVALSI	51.34%	—	—	—
Composite indicators				
Ind 1, school climate and human relations	0.96	0.09	1	0
Ind 2, managerial practices and principal's strategy	0.86	0.11	0.83	0.12
Ind 3, structures and resources of the school	0.94	0.09	1	0.11

3.2. Semiparametric expectation–maximization algorithm applied to Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione data

The aim of this subsection is to apply the EM algorithm for semiparametric mixed effects models to the INVALSI database for 2013–2014 as a tool for clustering Italian schools on the basis of their student attainments. The correlation between previous student scores (grade 6) and current student scores (grade 8) changes across schools, in that the effects that schools give to student attainments are heterogeneous and depend on different school characteristics. From this perspective, student scores at grade 8 can be seen as the result of student scores 2 years before (grade 6) combined with the effect of having attended a particular school for 2 years. The idea is to find out how student test scores at grade 6 and grade 8 are related to each other in different schools and in which schools these relationships are similar. In other words, we look for how many and which different trends exist in the scores of students attending Italian schools and, on the basis of the results, we group schools into different subpopulations. For this, the semiparametric EM algorithm works as an in-built classifier, since it groups schools into subpopulations, without knowing *a priori* the number of subpopulations.

On the basis of previous literature, it is reasonable to think that there is a linear correlation between student scores at grade 6 and at grade 8 (Agasisti *et al*., 2017; Masci *et al*., 2016, 2017). We therefore consider a semiparametric two-level linear model (where students represent the first level and schools the second), with student test scores at grade 6 and the student socioeconomic index as random and fixed effects respectively, allowing both the intercept and the coefficient of student test scores at grade 6 to be random or school specific. For each student j, $j = 1, \ldots, n_i$, and each school i, $i = 1, \ldots, N$, given that N is the total number of schools, J is the

Table 11. Maximum likelihood estimates of coefficients of model (21) obtained by applying the semiparametric EM algorithm to a selection of INVALSI data for 2013–2014

Subpopulation	$\hat{\beta}$	$\hat{c}_0$	$\hat{c}_1$	$\hat{w}$ (%)
1	1.417	46.028	0.454	12.2
2	1.417	22.579	0.707	39.6
3	1.417	30.293	0.648	37.5
4	1.417	31.207	0.393	8.8
5	1.417	25.359	0.027	1.9

total number of students and $\Sigma_{i=1}^{N} n_i = J$, the model takes the form

$$\begin{gathered} \mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{1}b_{0i} + \mathbf{z}_i b_{1i} + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N, \\ \boldsymbol{\epsilon}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{1}_{n_i}) \end{gathered} \tag{21}$$

where the answer variable $\mathbf{y}_i = (y_{1i}, \ldots, y_{n_i i})$ is the mathematics test score at grade 8, MATH8, of the n_i students within school i, whereas the covariate $\mathbf{z}_i = (z_{1i}, \ldots, z_{n_i i})$ and the covariate $\mathbf{x}_i = (x_{1i}, \ldots, x_{n_i i})$ are respectively the mathematics test score at grade 6, MATH6, and the socio-economic index ESCS of the n_i students within the ith school. The choice of considering ESCS as the fixed effect and MATH6 as the random effect is because we are interested in exploring how the correlation between MATH6 and MATH8, seen as a reflection of schools' ability in training students to achieve certain results, given their students' starting potential, varies between schools.

To have robust estimates, we select, from the data set that was presented in Section 3.1, only the schools that have at least 10 students. The resulting data set consists of 6188 students nested within 363 schools.

The semiparametric EM algorithm is applied, considering $\tilde{w} = 0.015$, $D = 0.8$, it $= 30$, itmax $=$ it1 $= 20$ and tollR $=$ tollF $= 10^{-4}$. Given these parameters, the algorithm identifies $M = 5$ distinct subpopulations, whose estimates of parameters are shown in Table 11.

The coefficient β in Table 11 is the coefficient related to ESCS (the fixed effect). Its positive value (1.417) suggests that, on average, students with a high socio-economic index are associated with high performances, in line with previous literature (Sirin, 2005). The estimated $\hat{w}_l$, for $l = 1, \ldots, M$, express the percentage of Italian schools belonging to each subpopulation l, for $l = 1, \ldots, M$. We identify two main subpopulations (subpopulation 2 and subpopulation 3 in Table 11) that contain about 77% of the total population, whereas the remaining 23% are distributed across the three other subpopulations. Regarding the analysis of the random-effects coefficients, Fig. 6 helps us to visualize them.

Looking at Fig. 6, it is immediately evident that there is quite an anomalous subpopulation, identified by the lowest regression plane, characterized by a very low slope (subpopulation 5 in Table 11). From an interpretative point of view, this subpopulation contains the 'worst' set of Italian schools. Indeed, it is characterized by both low intercept and low slope and this means that students in this kind of schools have on average low results at grade 8, even if they had good results at grade 6. In other words, students have on average low scores, without variability depending on their previous performances: students who had good results at grade 6, after attending 2 years in a secondary school belonging to subpopulation 5, have on average low performances, similarly to those of students who performed worse than them 2 years before. In

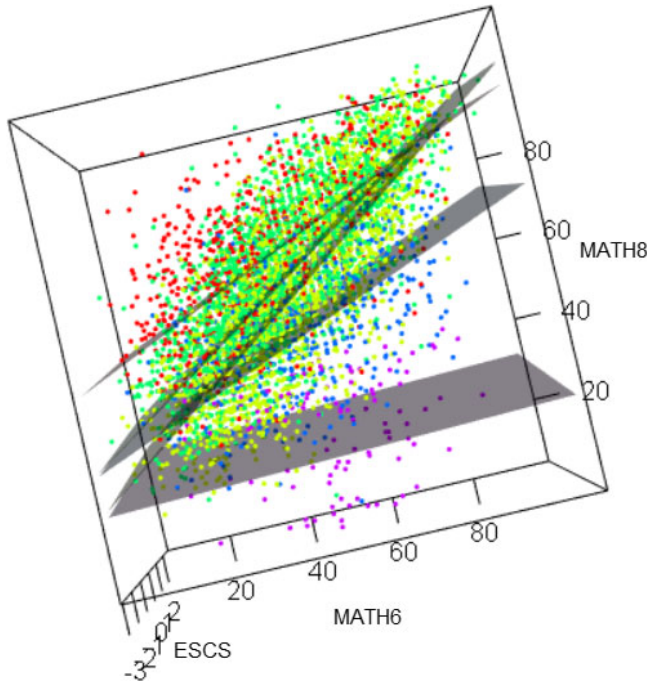


Fig. 6. Plot of the INVALSI data with the five regression planes identified by the semiparametric EM algorithm, for model (3): the parameters are shown in Table 11; colours represent the five subpopulations

contrast, the best scenario is represented by the subpopulation at the top of Fig. 6, identified by the regression plane with (subpopulation 1 in Table 11) the highest intercept (46.028) and a still high slope (0.454). These values suggest that even students who had very low scores at grade 6 obtain high scores at grade 8 with respect to their counterparts attending schools belonging to other subpopulations. Moreover, the value of the slope suggests that, even if students had on average an improvement in their performances, there is still heterogeneity across students who performed differently 2 years before, in that the best students continue to perform best compared with the average. It is worth noting that we are assuming homogeneity of variance across subpopulations. The variances of the errors in the five subpopulations are as follows: 128.33 for subpopulation 1, 103.07 for subpopulation 2, 97.92 for subpopulation 3, 133.36 for subpopulation 4 and 147.93 for subpopulation 5. By looking at the distributions of the errors across the five subpopulations, we do not have statistical evidence to reject the assumption of homogeneity of variance across subpopulations.

Thanks to the multilevel structure, we can also compute the percentage of variability explained by random effects, PVRE, which, in our case, is the percentage of variability in student test scores explained at school level:

$$\text{PVRE}_{\text{School}} = \frac{\sigma^2_{\text{School}}}{\sigma^2_{\text{School}} + \sigma^2_{\text{Residuals}}}.$$

Given the two-level semiparametric model

$$\mathbf{y}_i = \beta \mathbf{x}_i + c_{0l} + c_{1l}\mathbf{z}_i + \boldsymbol{\epsilon}_i;$$

Table 12. MSE computed by three models: a parametric fixed effects model, a parametric mixed effects models with both intercept and covariate as random effects and a semiparametric mixed effects model with both intercept and slope as random effects

	Parametric fixed effects model	*Parametric random-effects model*	*Semiparametric EM random intercept/slope*
MSE	155.91	111.55	118.69

the variance of the random effects is given by

$$\sigma^2_{\text{School}} = \sigma^2_{c_0} + 2\,\text{cov}(c_0, c_1)\bar{z} + \sigma^2_{c_1}\bar{z}^2.$$

Computing the empirical values of $\sigma^2_{c_0}$, $\text{cov}(c_0, c_1)$ and $\sigma^2_{c_1}$ from the estimated parameters, we obtain PVRE = 70.48%. This quantity confirms the significance of the random effects in explaining the result, since about 70% of the explained variability at student level is explained by differences across schools.

To provide an index for the goodness of fit of the model, we perform a leave-one-out cross-validation, we compute the mean-square error (MSE) and we compare it with those obtained considering

(a) the same model but with all the parameters as fixed effects and
(b) the parametric mixed effects models with the same choice of random and fixed effects.

Table 12 reports the three MSEs computed on the student test scores.

The MSE that was obtained with the fixed effects model is the highest (155.91) and it departs from those obtained by both the parametric and the semiparametric mixed effects models (111.55 and 118.69 respectively). On the basis of the nature of the problem, we expect the parametric mixed effects model to perform the best, since it fits the trend of the data within each school. Nonetheless, the semiparametric mixed effects model produces a slightly bigger MSE, but it extracts a new kind of information from the data. Indeed, whereas the parametric approach can estimate the parameters of a model, i.e. based on an already known structure of the data, the semiparametric approach takes a further step, since it can identify a new structure within the data, i.e. the existence of a new latent level of grouping. From an interpretative point of view, the identification of subpopulations is highly informative in identifying those groups that depart from the usual behaviour. Indeed, among the identification of subpopulations itself, what really matters is the identification of the minority subpopulations, which are those subpopulations containing a small percentage of the entire population, characterized by different properties with respect to the majority. In our application to the INVALSI database, subpopulations 2 and 3, which are very close to each other and contain almost 80% of the schools, represent the most common trend, but the subpopulations that deserve more attention are subpopulations 1, 4 and 5, which are those containing a smaller percentage of schools that behave differently from the majority. Moreover, the relatively small difference between the MSEs of the two approaches suggests that the subpopulations structure that is identified by the semiparametric EM algorithm catches almost all the heterogeneity across the effects of Italian schools, meaning that the subpopulations are quite homogeneous.

A further consequence of the identification of a latent structure within the data is that subpopulations are likely to derive from some unknown characteristics of schools, which lead

to these differences. In general, the interpretation a *posteriori* of subpopulations of data is important *per se*, especially in terms of 'big data', where the identification of patterns within a large amount of data, marked by a complex and unknown structure, is particularly relevant. For this reason, in the next subsection, we try to find out whether there are patterns of school level variables that characterize the estimated subpopulations.

3.3. Association between school characteristics and school subpopulations

Applying the semiparametric EM algorithm to the INVALSI data, we discover a structure of subpopulations that clearly reflects heterogeneities among the effects of Italian schools. In particular, we identify five different subpopulations, that emerge from five different behaviours of schools in affecting the evolution of their student achievements. We are interested in exploring these subpopulations *a posteriori*, to investigate whether there are school characteristics that are associated with them. Among these five subpopulations, subpopulation 2 and subpopulation 3 in Table 11, which are characterized by similar parameters and which contain almost 80% of the entire set of schools, represent the majority of schools. Consequently, we consider the union

Table 13. Results of the multinomial logit model (22)†

Variable name	*Result for subpopulation 1*	*Result for subpopulation 4*	*Result for subpopulation 5*
Intercept	−2.287	−0.560	−23.363
Mean ESCS	−0.335	−0.043	0.087
Female percentage	0.013	−0.016	−0.011
Immigrant percentage	−0.069	−0.077	−0.246
Early enrolled student percentage	−0.095	0.030	0.014
Late enrolled student percentage	0.013	0.034	−0.012
Number of classes	−0.035	−0.008	0.067
Number of school complexes	0.078	−0.126	0.086
Private	0.884	−9.187‡	−6.147‡
Principal features			
Gender (female ≡ 1)	−0.192	−0.043	0.211
Age (years)	0.018	−0.048	0.020
Master after degree (yes ≡ 1)	0.478	−0.577	0.981
Scientific education (yes ≡ 1)	−0.135	0.171	−6.019‡
Years of experience	0.013	0.035	0.046
Years of experience in the actual school	−0.096	−0.034	−0.048
Experience in another district	0.004	0.583	−1.390
Experience with INVALSI	−0.155	0.384	1.525
Composite indicators			
Ind 1: school environment and human relations	0.327	−0.083	1.864
Ind 2: managerial practices and principal's strategy	2.899	−0.626	−5.762
Ind 3: structures and resources of the school	−3.588	2.726	6.553
Geographical area			
Centre	0.744	0.648	15.691‡
South	1.201	1.200	14.687‡

†Coefficients are computed considering S_{ref}, the union of subpopulations 2 and 3, as the reference.
‡p-value less than 0.0001.

of subpopulations 2 and 3 as the reference subpopulation S_{ref}, which represents the reference trend. Our interest is to see how the school characteristics of the other three subpopulations (subpopulations 1, 4 and 5 in Table 11) differ from the reference subpopulation. For this, we apply a multinomial logit model by treating all the school level characteristics that are shown in Table 10 as covariates and, as outcome variable, belonging to the four subpopulations.

For each group (school) $i = 1, \ldots, N$ and each subpopulation $l = \{1, 4, 5\}$, the model takes the form

$$\ln\left\{\frac{P(Y_i = l)}{P(Y_i = S_{\text{ref}})}\right\} = \beta_{0l} + \sum_{q=1}^{Q} \beta_{lq} X_{iq}, \tag{22}$$

where X is the $N \times Q$ matrix of school level covariates shown in Table 10, with Q the total number of school level covariates. The results of model (22) are shown in Table 13.

Among the many school level variables, only four variables turn out to be associated with schools' belonging to the four subpopulations: the percentage of immigrants, the dummy for public or private school, the kind of education of the school principal (humanities or scientific) and the geographical area (northern, central and southern Italy). With respect to the reference subpopulation, subpopulations 1 and 4 are more likely to contain schools with low percentages of immigrant students; clusters 4 and 5 are less likely to contain private schools; subpopulation 5 is more likely to contain schools that are managed by school principals with a humanities rather than a scientific education; subpopulations 1 and 4 are more likely to contain schools in southern Italy and subpopulation 5 is most likely to contain schools in both central and southern Italy. The fact that subpopulations 1 and 4 are more likely to contain schools with low percentages of immigrant students and are also more likely to contain schools in southern Italy is an expected result since the majority of immigrant students in Italy live in northern Italy. Subpopulations 1 and 4 are also the subpopulations with the highest intercepts and high positive slopes (see Table 11), being the best scenario of schools on the basis of our interpretation, and those schools turn out to be associated with southern Italy and to low percentages of immigrant students. The fact that both subpopulations 4 and 5 are less likely contain private schools reveals that private schools tend to be associated neither with the worst set of schools (subpopulation 5 of Table 11) nor to a very good set of schools (subpopulation 4 of Table 11).

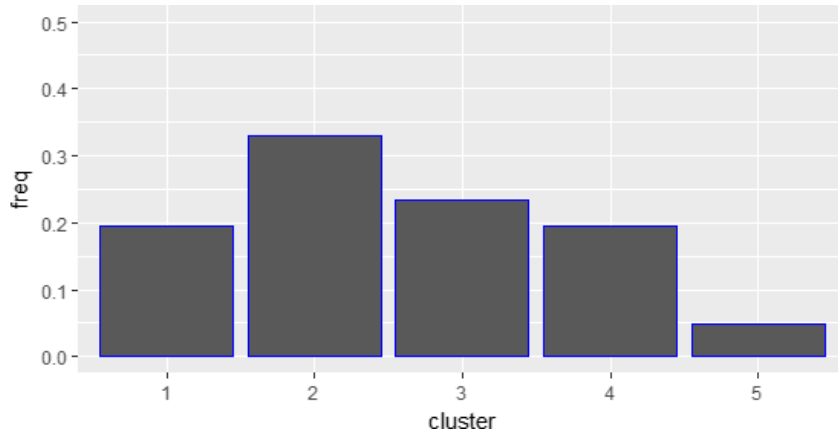
Geographical differences represent an interesting aspect in the Italian educational context. Fig. 7 reports the proportion of schools belonging to the five subpopulations, in the three geographical Italian macroareas: northern, central and southern Italy.

Comparing northern and southern Italy, we can see that the distribution of schools among subpopulations is different. In northern Italy, we do not have any schools belonging to subpopulation 5 and we have very few schools belonging to subpopulations 1 and 4: almost all schools belong to subpopulations 2 and 3. In southern Italy, the distribution of schools among subpopulations is more uniform and it is possible to count a good quantity of schools belonging to each subpopulation.

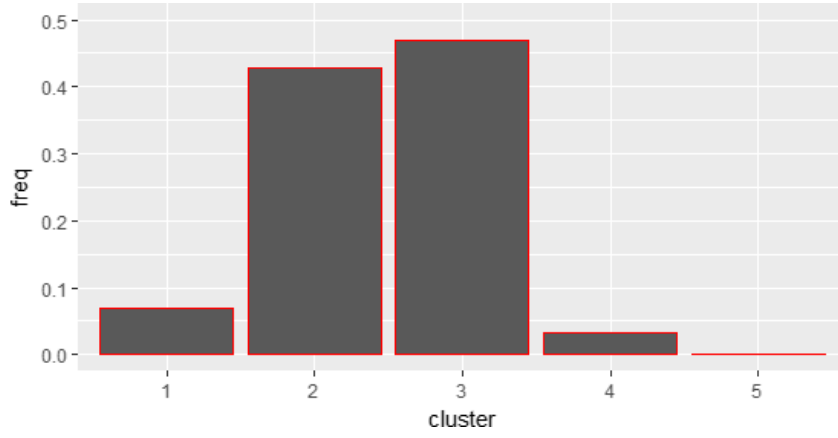
The fact that, among the entire set of school level variables at our disposal, only four variables turn out to be significantly associated with the presence of subpopulations does not imply that there is no explanation for the presence of subpopulations of schools, but, most likely, these subpopulations derive from other dynamics, that we cannot observe or measure.

4. Conclusions

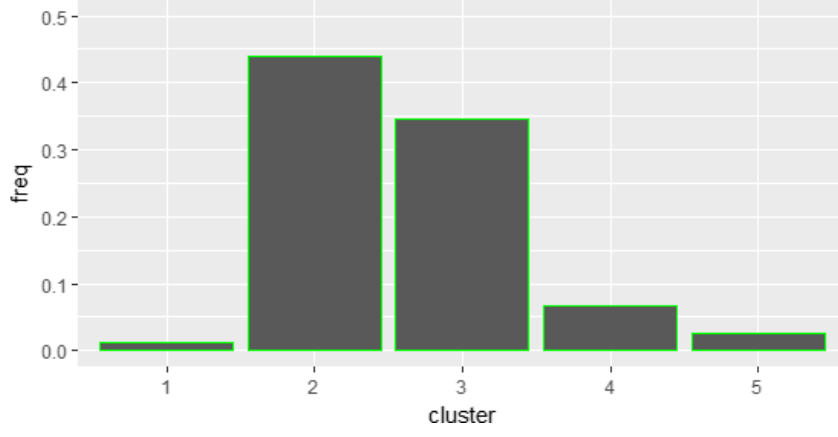
This paper proposes an EM algorithm for semiparametric mixed effects models (the semiparametric EM algorithm), presents a simulation study and applies the semiparametric EM



(a)



(b)



(c)

Fig. 7. Proportion of schools belonging to the five subpopulations, within the three geographical Italian macroareas (a) northern, (b) central and (c) southern Italy

algorithm to INVALSI data for 2013–2014 as a tool for clustering Italian schools. The semiparametric EM algorithm places itself in the branch of literature concerning the algorithms proposed in Aitkin (1996) and Azzimonti *et al*. (2013). In particular, our algorithm is inspired by that proposed in Azzimonti *et al*. (2013), but it introduces the major improvement, among others, that the covariates are group specific, meaning that they can vary both in number of observations and in range of assumed values across groups. Moreover, with respect to the algorithm proposed in Aitkin (1996) and the literature about GMMs and latent class analysis, the advantage of the semiparametric EM algorithm is that it does not need to fix *a priori* the number of discrete masses (subpopulations), but, conditionally on certain parameter values, the algorithm itself identifies the number of discrete support points. This has great value in applications where the number of subpopulations is not known *a priori* and the aim is therefore to find out how many and which different trends exist within the data. This concept is particularly relevant in the era of big data, where there is the need to identify latent structures within big and complex databases.

The semiparametric EM algorithm, when applied to the INVALSI data, can identify subpopulations of schools, within which student achievement trends differ. Among the identification of the number of subpopulations, which reveals how many different trends exist within the sample of Italian schools, the weights that are associated with the subpopulations give further information about the clustering. In a context in which we do not know *a priori* which is the expected trend, the subpopulations that are associated with higher weights represent the most common behaviour, whereas the less numerous subpopulations (those associated with lower weights) represent those schools whose impact differs from the majority. This draws attention to what determines whether schools belong to the minority subpopulations. In particular, the algorithm identifies five school subpopulations that represent different school associations with their student achievements trends, seen as the ability of junior secondary schools to train students to obtain certain skills at the end of the 3 years, given their skills at the beginning of schooling, adjusting for their socio-economic index ESCS. In the INVALSI framework, schools are associated with a *positive or negative effect*, based on the final performances of their students and given their students' initial skills. Among these five subpopulations, a subpopulation containing schools with a negative effects is immediately evident. This subpopulation contains schools that have students who tend to underperform, with respect to their performance 2 years before, since they have on average very low scores, even if 2 years before, when they started to attend these schools, they obtained higher scores. Regarding positive effects, we interpret the subpopulation with the highest intercept and positive slope (subpopulation 1) as the best, in terms of school effect, since it contains schools that can train students to reach high performances, even if they had low performances at the beginning of schooling. It is worth saying that, from a policy perspective, the definition of the *best school effect* is currently in debate. Indeed, it is reasonable to consider a school in which all students obtain very high scores, without heterogeneity, as a school with a good effect, but, in contrast, a different point of view emphasizes the advantages of having heterogeneity within the school. In this perspective, the role of the school is continuously to raise the students goals to urge pupils to perform even better, using competition and variation to motivate them.

After the identification of school subpopulations, the paper focuses on another actual and interesting topic, i.e. their interpretation *a posteriori*. In particular, we explore the associations between school subpopulations and school level characteristics, showing that only geographical areas, the percentage of immigrants, a dummy for private or public school and school principal's education turn out to be significantly associated. This evidence suggests that the school level variables at our disposal do not explain the differences in school effects. On the basis of the

fact that the school subpopulations are clearly different in their effect on student attainments, the lack of stratification of school level variables across subpopulations might mean that the observed school level variables do not reflect the real school characteristics (i.e. they are not measured in the right way) or there are other latent aspects, that we cannot measure, which might explain the different effects of schools on their students.

In the future, our aim is to deepen the analysis on the characterization of the estimated school subpopulations, considering other information about the school environment, which we have not been able to measure until now. Moreover, from a methodological point of view, our aim is to develop a multivariate version of the EM algorithm for semiparametric mixed effects models, to consider two (or more) response variables and to relax the linearity assumptions, considering also the case of other functional forms. In the INVALSI framework, since the data set contains both student scores in reading and in mathematics, it would be possible to apply the multivariate version, in which the response variable would be the bivariate vector of reading and mathematics scores, and, consequently, to cluster schools or classes on the basis of both their effects on reading and mathematics student attainments, analysing the interactions between these two fields.

References

Agasisti, T., Ieva, F. and Paganoni, A. M. (2017) Heterogeneity, school-effects and the north/south achievement gap in Italian secondary education: evidence from a three-level mixed model. *Statist. Meth. Appl.*, **26**, 157–180.

Agasisti, T. and Vittadini, G. (2012) Regional economic disparities as determinants of student's achievement in Italy. *Res. Appl. Econ.*, **4**, no. 2.

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statist. Comput.*, **6**, 251–262.

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

Bock, R. D. (2014) *Multilevel Analysis of Educational Data*. Amsterdam: Elsevier.

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

Bryk, A. S. and Raudenbush, S. W. (1988) Toward a more appropriate conceptualization of research on school effects: a three-level hierarchical linear model. *Am. J. Educ.*, **97**, 65–108.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F. and York, R. (1966) The Coleman report. *Equality of Educational Opportunity*. Washington DC: US Government Printing Office.

Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J. R. Statist. Soc.* A, **159**, 385–443.

Hanushek, E. A., Rivkin, S. G. and Taylor, L. L. (1996) Aggregation and the estimated effects of school resources. *Technical Report*. National Bureau of Economic Research, Cambridge.

Heinen, T. (1996) *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. New York: Sage.

Lin, L. I. (2000) Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statist. Med.*, **19**, 255–270.

Lindsay, B. G. (1983a) The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, **11**, 86–94.

Lindsay, B. G. (1983b) The geometry of mixture likelihoods, part ii: the exponential family. *Ann. Statist.*, **11**, 783–792.

Masci, C., De Witte, K. and Agasisti, T. (2018) The influence of school size, principal characteristics and school management practices on educational performance: an efficiency analysis of Italian students attending middle schools. *Socio-Econ. Planng Sci.*, **61**, 52–69.

Masci, C., Ieva, F., Agasisti, T. and Paganoni, A. M. (2016) Does class matter more than school?: Evidence from a multilevel statistical analysis on Italian junior secondary school students. *Socio-Econ. Planng Sci.*, **54**, 47–57.

Masci, C., Ieva, F., Agasisti, T. and Paganoni, A. M. (2017) Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. *J. Appl. Statist.*, **44**, 1296–1317.

Masci, C., Johnes, G. and Agasisti, T. (2019) Student and school performance across countries: a machine learning approach. *Eur. J. Oper. Res.*, to be published.

McCulloch, C., Lin, H., Slate, E. and Turnbull, B. (2002) Discovering subpopulation structure with latent class mixed models. *Statist. Med.*, **21**, 417–429.

Muthén, B. (2004) Latent variable analysis. In *The Sage Handbook of Quantitative Methodology for the Social Sciences*, pp. 345–368. Thousand Oaks: Sage.

Muthén, B. and Shedden, K. (1999) Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55**, 463–469.
Nagin, D. S. (1999) Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychol. Meth.*, **4**, no. 2, 139–157.
Pinheiro, J. C. and Bates, D. M. (2000) Linear mixed-effects models: basic concepts and examples. In *Mixed-effects Models in S and S-Plus*, pp. 3–56. New York: Springer.
Proust-Lima, C., Letenneur, L. and Jacqmin-Gadda, H. (2007) A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome. *Statist. Med.*, **26**, 2229–2245.
Raudenbush, S. and Bryk, A. S. (1986) A hierarchical model for studying school effects. *Sociol. Educ.*, **59**, 1–17.
Raudenbush, S. W. and Willms, J. (1995) The estimation of school effects. *J. Educ. Behav. Statist.*, **20**, 307–335.
R Development Core Team. (2014) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Sani, C. and Grilli, L. (2011) Differential variability of test scores among schools: a multilevel analysis of the fifth-grade INVALSI test using heteroscedastic random effects. *J. Appl. Quant. Meth.,* **6**, 88–99.
Sarrico, C. S., Rosa, M. J. and Manatos, M. J. (2012) School performance management practices and school achievement. *Int. J. Product. Perform. Mangmnt*, **61**, 272–289.
Sirin, S. R. (2005) Socioeconomic status and academic achievement: a meta-analytic review of research. *Rev. Educ. Res.*, **75**, 417–453.
Vanthienen, J. and De Witte, K. (2017) *Data Analytics Applications in Education*. New York: Taylor and Francis.
Vermunt, J. K. and Magidson, J. (2002) Latent class cluster analysis. *Appl. Latnt Class Anal.*, **11**, 89–106.