MOX-Report No. 63/2017

Non-parametric mixed-effects models for unsupervised classification of Italian schools

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

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

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

Non-parametric mixed-effects models for unsupervised classification of Italian schools

Chiara Masci[♮], Anna Maria Paganoni[♮], Francesca Ieva[♮]

November 22, 2017

[♮] MOX - Modelling and Scientific Computing, Department of Mathematics,
Politecnico di Milano, via Bonardi 9, Milano, Italy
chiara.masci@polimi.it
anna.paganoni@polimi.it
ieva.francesca@polimi.it

Abstract

This paper proposes an EM algorithm for non-parametric mixed-effects models (NPEM algorithm) and applies it to the National Institute for the Educational Evaluation of Instruction and Training (INVALSI) data of 2013/2014 as a tool for unsupervised clustering of Italian schools. The main novelties introduced by NPEM algorithm, when applied to hierarchical data, are twofold: first NPEM allows the covariates to be group specific; second, it assumes the random effects to be distributed according to a discrete distribution P^* with an (a priori) unknown number of support points. In doing so, it induces an automatic clustering of the grouping factor at higher level of hierarchy, enabling the identification of latent groups of schools that differ in their effect on student achievements. The clustering may then be exploited through the use of school level features.

Keywords: EM algorithm; Non-parametric mixed-effects models; Student achievements; School value-added.

1 Introduction

The analysis of education systems is a subject receiving particular attention in the last decades. During their learning process, students are influenced by multiple aspects coming from both their personal and school life. Personal motivation, family, friends and geographical context play a fundamental role in education student performance and the choice of the school is also particularly relevant. The literature provides numerous studies aimed at measuring and explaining the "school effect", intended as the impact that the school the student is attending has on his/her achievements, [6], [8], [9] and [18]. In [6], the authors state the importance of considering the "unit-of-analysis" (students, classes, schools), when speaking about educational research, and they argue that hierarchical models should constitute the basic paradigm for quantitative research on student learning. Also, in [18], the authors, given the hierarchical structure of education data, underlie the importance of measuring school effects and present different approaches to analyze nested data. In the Coleman report [8], the author views the education as a process in which students' performance (output) is produced from inputs including school resources, teacher quality, family attributes, and peer quality. In his perspective, policy attention should be focused on inputs that are both directly controlled by policymakers (characteristics of schools, teachers, curricula, etc.) and those that are "uncontrolled" (family, friends, the learning capacities of the student, etc.). Also Hanushek, in [9] shows that schools' characteristics are of importance in determining student outcomes.

The nature and the magnitude of the school impact on students attainments strongly depend on the type of school system and related regulations. There are countries where the education system is totally centralized and, therefore, school programs and practices are very homogeneous across the territory. On the other hand, in the last years the dynamics of education systems are changing and more and more countries are decentralizing the power on decision about education, giving more autonomy to schools [20]. This phenomenum leads to differences across schools that are reflected on differences across student achievements. The Programme for International Student Assessment (OECD-PISA, `www.pisa.oecd.org`) tests 15 year-old students in mathematics, reading and science in more than 70 countries all over the world, every three years since 2000. Studies on PISA data show that Italy is a country where the percentage of variability in student achievements given to the grouping factor within schools is quite high with respect to other countries [10]. This means that in Italy the added-value that schools give to their students is relevant: in other words, attending a certain school instead of another produces an effect on student's skills. Schools differ under many aspects: size, location, school body composition, teachers, school principal management style and much more. 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 (INVALSI) tests students all over the country since 2004, both in their mathematics and literature skills, following a procedure similar to OECD-PISA. These tests are done 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 confirm that the magnitude of the school effect on student attainments is substantial. In [1], [13] and [14] the authors observe that the percentage of variability in student attainments in INVALSI tests explained by the school (PVRE) depends on the geographical macro-area 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 of PISA data in Italy report that, in mathematics, the PVRE exceeds the 40% [10].

An important characteristic of educational administrative data is their hierarchical structure: stu-

dents are naturally nested within schools. In the perspective of the learning process investigation, it is important to disentangle the effects 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, [5], [22]. Indeed, multilevel models take into account the hierarchical nature of data and are able to quantify the part of variability in the response variable that is given to each level of grouping, [16]. In particular, in the case of students nested within schools, they are able to estimate the "school effect", that is the value-added of the school to its student achievements. In the context of education research, the use of fixed (FE) and random effects (RE) in hierarchical regression is frequently discussed and, in the last decades, their effectiveness in terms of policy-relevant inference has been carefully analyzed. In [7], the authors compare the robustness of FE models with the modelling flexibility and potential efficiency of RE models, in a two-level hierarchical linear regression. The common issue concerning both the two approaches is linked to what economics literature calls "exogeneity" assumption (i.e. assuming that the individual-level residuals are independent of the covariates), that in a policy-relevant perspective is crucial in order to interpret the estimated coefficients as causal effects, while, when the assumption does not hold, they can be considered as estimates of associations. Moreover, RE models add the so called "RE assumption", i.e., that random effects are uncorrelated with any of the covariates used in the model. Again, this assumption is crucial when the intention is to use RE models for policy causal inference. This is why, in the education economic context, it is important to adjust the model for unobserved characteristics, at student, family and school levels. Also, in [15] the authors state that the unobserved characteristics at both student and school levels are one of the main issues that bias the estimates of value-added education production models, when the aim is to explain the effect of school inputs and past skills on student test scores. In this perspective, it is worth to notice that the aim of this work is to identify latent clusters of schools that differ in the association of their student attainments across different years. Therefore, we model these clusters by choosing as random effects a discrete distribution P^* with an unknown finite number of mass points, that is able to detect a latent structure among the Italian schools.

From a practical point of view, in Italy students must attend five years of primary school, three years of junior secondary school and five years of upper secondary school. If we focus on junior secondary schools, the "school effect" can be seen as the ability of these schools in receiving students from the primary schools with certain skills and give them new and increased skills at the end of the three years. Our analysis aims then at identifying clusters of schools, standing on the relationship between their students test scores at the beginning and at the end of the three 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 students 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. The scope is to identify clusters of schools within which schools perform in a similar way (in the sense that the linear relation between their students scores at grades 6 and 8 is similar) and across which they perform differently.

In the literature, multilevel linear models have already been applied to INVALSI data, with a view to estimating schools value-added, modeled by means of parametric distributions, after adjusting for students characteristics, [1], [13], [14] and [19]. The method that we present in this article is new and has a different scope with respect to the previous literature. Our aim is to develop and apply an EM algorithm for non-parametric mixed-effects models (NPEM algorithm) for hierarchical data (students nested within schools), in order to perform an in-built classifier of the grouping factor (schools). 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 non-parametric and it follows a discrete distribution with an

unknown number of masses. The algorithm itself identifies the number of masses, not fixed a priori, that represent clusters in which schools are grouped, standing on the achievements trend of their students. Both the algorithm and its application to the educational context are new to the literature.

The consequence of the identification of clusters of schools is that we can recognize how many and which different behaviors characterize Italian schools and, therefore, identify a latent structure within them. In a second stage, this enables the profiling of clusters by means of school level variables. The idea is that there could be variables at school level that influence the school effect on student achievements. Therefore, in the second part of the analysis we explore the presence of patterns of school characteristics among clusters of schools.

The paper is organized as follows: in Section 2 we describe the model and methods - NPEM algorithm - together with a simulation study; in Section 3 we present the INVALSI dataset and report the application of NPEM algorithm to INVALSI data, shows the results and explores the relation between clusters and school characteristics; in Section 4 we draw our conclusions.

All the analysis are made using R software [17]. The code for NPEM algorithm is available upon request to the authors.

2 Model, methods and simulation study

In this section, we present the non-parametric 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).

2.1 Non-parametric mixed-effects model

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

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

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

In the parametric framework of mixed-effects linear models, coefficients of random effects 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 [16].

The main novelty introduced here is that we move to a non-parametric framework, assuming the coefficients $\mathbf{b}_i$ to be distributed according to a discrete distribution P^*, assuming M sets of values $(c_{0l}, \dots, c_{rl})$ for $l = 1, \dots, M$, where $M \leq N$. This means that each group i, for $i = 1, \dots, N$, is assigned to a cluster l, that is characterized by random parameters $(c_{0l}, \dots, c_{rl})$. This non-parametric modelling enables 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, first of all,

we can identify how many latent clusters exist within the groups of data and, second, we can estimate the parameters associated to each cluster, pointing out their differences.

Under these assumptions, the non-parametric mixed-effects model takes the following form:

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

In particular, from now on, without loss of generality, we consider the case with one random intercept, one random effect and one fixed effect[1]:

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

where $\mathbf{1}$ is the n_i-dimensional vector of 1, $M \leq N$ is the number of clusters (mass points) unknown a priori. Coefficients $\mathbf{c}_l$, for $l = 1, \dots, M$, 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 in (3). The ML estimator $\hat{\mathcal{P}}^*$ of $\mathcal{P}^*$ can be obtained following the theory of mixture likelihoods in [11] and [12], where the author proves the existence, discreteness and uniqueness of the non-parametric maximum likelihood estimator of a mixing distribution, in the case of exponential family densities. In particular, the author faces 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 ML estimator of the random effects distribution can be expressed as a set of points $(\mathbf{c}_1, \dots, \mathbf{c}_M)$, where $M \leq N$ and $\mathbf{c}_l \in \mathbb{R}^2$ for $l = 1, \dots, M$, and a set of weights $(w_1, \dots, w_M)$, where $\sum_{l=1}^{M} w_l = 1$ and $w_l \geq 0$ for each $l = 1, \dots, M$. Given this, we propose an algorithm for the joint estimation of σ^2, β, $(\mathbf{c}_1, \dots, \mathbf{c}_M)$ and $(w_1, \dots, w_M)$, that is performed through the maximization of the likelihood, mixture by the discrete distribution of the random effects,

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

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

The algorithm that we propose is inspired by the one proposed in [4], but it considers the linear functional dependence between response and predictors and it makes three main improvements: (i) the optimization of the Maximization step is computed in closed form, (ii) the covariates can be group specific and (iii) the initialization of the parameters' range is computed in a more robust and generalizable way. The first point directly derives from the linearity assumption, while the second one means that, differently from [4], we allow the group covariate to be different both in number of observations and in the values they assume across groups. Regarding the initialization of the parameters' range, we define it computing a single regression for each one of the N groups and assuming a uniform distribution of N mass points among the estimated parameters. The idea at the base of the

[1] This choice is due to the case considered in the application to INVALSI dataset, in Setion 3.2

algorithm is also similar to the one proposed in [3], but while in [3] the authors need to fix a priori the number of discrete points of the mixing distribution, our algorithm identifies itself the number of support points M, standing on given tolerance values that we fix depending on the problem.

2.2 The NPEM algorithm

The proposed EM algorithm is an iterative algorithm that alternates two steps: the expectation step (E step) in which we compute the conditional expectation of the likelihood function with respect to the random effects, given the observations and the parameters computed in the previous iteration; and the maximization step (M step) in which we maximize the conditional expectation of the likelihood function. At each iteration, the EM algorithm updates the parameters in order to increase the likelihood in (4) and it continues until a fixed number of iterations (`IT`) is reached. In particular, the update is given by:

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

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

where

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

and

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

Coefficients W_{il} represent the probability of $\mathbf{b}_i$ being equal to $\mathbf{c}_l$ conditionally to observations $\mathbf{y}_i$ and given the fixed coefficient β and the variance σ^2.

The maximization in equation (6) involves two steps and it is done iteratively. In the first step, we compute the *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 with respect to all support points $\mathbf{c}_l$ separately, that means

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

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 coefficients of random effects are:

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

and

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

In the second step, we fix the support points of the random effects distribution computed in the previous step and we compute the *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), are given by:

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

and

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

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

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

Therefore, in order 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. So that, the estimation of the coefficients $\mathbf{b}_i$ of the random effects for each group is obtained maximizing W_{il} over l, that is

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

The algorithm starts considering a given distribution with N support points for the coefficients of random effects and a starting estimate for the coefficients of fixed effects. In particular, the starting N support points are obtained fitting a simple regression within each group and estimating the couple of parameters (both the intercept and the slope) for each one of the N groups. The weights are uniformly distributed on these N support points (each weight is equal to $1/N$). Regarding the starting values of β and σ^2, they are estimated fitting a unique linear regression on the entire population.

During the iterations, the EM algorithm performs the support reduction of the discrete distribution, in order to identify $M < N$ mass points in which the N groups are clustered. The support reduction is

made standing on two criteria. The former 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 $\mathbf{c}_{l,k} = \frac{\mathbf{c}_l + \mathbf{c}_k}{2}$ with weight $w_{l,k} = w_l + w_k$. The latter is that, starting from a given iteration up to to end, we fix a threshold $\tilde{w}$ and we remove mass points with weight $w_l \leq \tilde{w}$ or that are not associated to any group. When one or more mass points are deleted, the remaining weights are reparameterized in such a way that they sum up to 1:

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

where M^{new} is the total number of masses after deleting the ones associated to weight $w_l \leq \tilde{w}$ or not associated to any group, $\mathbf{w}^{old}$ are the old remaining weights and $\mathbf{w}^{new}$ are the new reparameterized weights.

The sketch of the algorithm is shown in Algorithm 1. Regarding the estimates of the parameters of Eq. (3), the algorithm updates them until convergence or until it reaches the maximum number of iterations fixed a priori for this stage (`itmax`). The convergence is reached when all the differences between the estimates of the parameters at 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 (`tollF` and `tollR` respectively).

The choice of the maximum numbers of iterations at different steps (`it`, `it1`, `itmax`) depends on the complexity of the data and on the consequent velocity of their clustering. The thresholds D and $\tilde{w}$ are two complexity parameters that affect the estimation of the non-parametric distribution: D governs the minimum difference (in terms of distance) between clusters - the higher is D, the lower is the number of clusters; $\tilde{w}$ reflects the minimum percentage of groups that we allow within each cluster. The values of `tollF` and `tollR` depend on the scale of the parameters.

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 converges in less than 30 iterations.

Algorithm 1: EM algorithm for non-parametric 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, itmax.

output: Final estimates of $\mathbf{c}_l^{(it)}$, $w_l^{(it)}$, for $l = 1, \ldots, M$, $\beta^{(it)}$ and $\sigma^{2(it)}$.

for $k \in 1, \ldots, it$ **do**

compute the distance matrix DIST (where $DIST_{st} = \sqrt{(c_{0s} - c_{0t})^2 + (c_{1s} - c_{1t})^2}$ is the euclidean distance between each couple of mass points $s, t \; \forall s, t = 1, \ldots, M, s \neq t$);

if $DIST_{st} < D \quad (\forall s, t = 1, \ldots, M, s \neq t)$ **then**

collapse masses s and t to a unique mass point;

compute the new distance matrix DIST;

if $k \geq it1$ **then**

if $w_l^{(k)} \leq \tilde{w} \quad (\forall l = 1, \ldots, M)$ **then**

delete mass point l;

reparameterize the weights according to Eq. (16);

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 Eq. (7);

update the weights $w_1^{(k)}, \ldots, w_M^{(k)}$ according to Eq. (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 Eq. (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 Eq. (12) and (13);

j=1;

while $(|\beta^{(k,j-1)} - \beta^{(k,j)}| \geq tollF \quad or \quad |\sigma^{2(k,j-1)} - \sigma^{2(k,j)}| \geq tollF \quad or \quad |\mathbf{c}_l^{(k,j-1)} - \mathbf{c}_l^{(k,j)}| \geq tollR) \quad \& \quad j \leq itmax$ **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 Eq. (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 Eq. (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 cluster l for each group i according to Eq. (15);

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 INVALSI dataset, in which we make this choice of fixed and random parameters. Nonetheless, the NPEM algorithm allows 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

In order to validate the proposed estimation algorithm, we perform a simulation study, considering a linear model with two covariates. In particular, we test the algorithm on a simulated dataset containing 9 groups of variables, where each group is composed by an answer variable and two covariates. We sample the variables in order to have 3 different clusters within the 9 groups, that is, in order to create nine cohorts of data characterized by three different linear correlations. For this purpose, we generate 9 response variables as the result of 3 distinct linear combinations of 3 couples of covariates, plus some errors:

$$\begin{cases} \mathbf{y}_i = \beta \mathbf{x}_1 + c_{01} + c_{11}\mathbf{z}_1 + \boldsymbol{\epsilon}_i & i = 1, \dots, 3. \\ \mathbf{y}_i = \beta \mathbf{x}_2 + c_{02} + c_{12}\mathbf{z}_2 + \boldsymbol{\epsilon}_i & i = 4, \dots, 6. \\ \mathbf{y}_i = \beta \mathbf{x}_3 + c_{03} + c_{13}\mathbf{z}_3 + \boldsymbol{\epsilon}_i & i = 7, \dots, 9. \end{cases} \tag{17}$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, 0.01)$ and the covariates are sampled by Normal distributions with different parameters. In particular,

$$\begin{aligned} &\mathbf{x}_1 \sim \mathcal{N}(0.30, 0.16), \qquad &&\mathbf{z}_1 \sim \mathcal{N}(0.10, 0.16), \\ &\mathbf{x}_2 \sim \mathcal{N}(0.28, 0.16), \qquad &&\mathbf{z}_2 \sim \mathcal{N}(0.12, 0.16), \\ &\mathbf{x}_3 \sim \mathcal{N}(0.27, 0.16), \qquad &&\mathbf{z}_3 \sim \mathcal{N}(0.08, 0.16), \end{aligned} \tag{18}$$

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. The choice of the size, of the parameters and of the distribution is arbitrary. We choose values of x and z that are in the same range and sizes that are between 90 and 100 in order to ease the visualization of the data. Other choices are possible and do not affect the validity of results. Lastly, coefficients in Eq. (17) are reported in Table 1.

	c_0	c_1	β
l=1	5	10	3
l=2	2	5	3
l=3	0	2	3

Table 1: Coefficients used for data simulation in Eq. (17). Each row corresponds to a cluster l. The intercept and the coefficient of z differ across groups (c_0 and c_1 respectively), while the coefficient of x (β) is fixed.

Again, the choice of the coefficients is arbitrary. For coherence with the upcoming INVALSI case study, that considers both the slope and the intercept as random, we choose different values for both the intercept and the coefficient of variable z across the three clusters, while we maintain the coefficient of x fixed[2]. In particular, we assign groups i = {1,2,3} to cluster l = 1, groups i = {4,5,6} to cluster

[2] Again, we choose values of the parameters $\mathbf{c}_l$, for $l = 1, \dots, 3$, and β in the range between 1 and 10, in order to ease the visualization of the results.

l = 2 and groups i = {7,8,9} to cluster l = 3 and therefore response **y** is generated according to this choice. Figure 1 shows the 3d image of the simulated data.

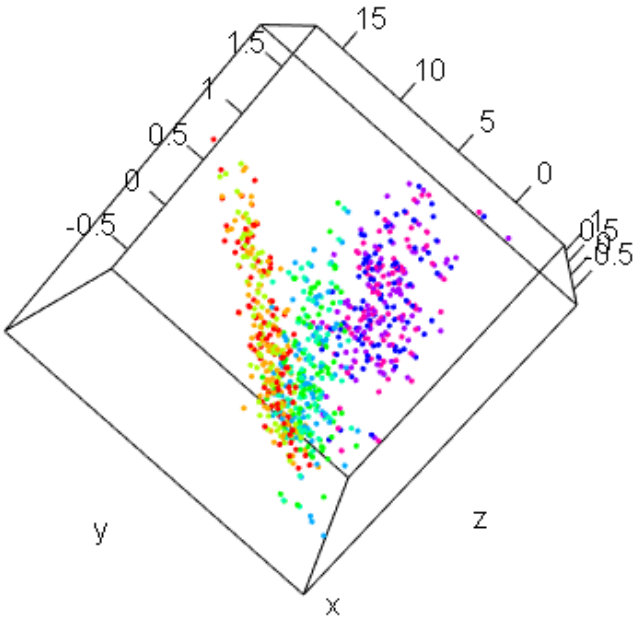

Figure 1: Plot of the simulated data obtained by Equations (18) and (17). Each one of the nine groups has a different color. Data with similar behaviors are automatically assigned to similar colors by software R.

Looking at the figure, it is possible to recognize three different linear correlations among the data, identified by the three distinct "clouds" of points. Groups of points characterized by similar linear correlations are automatically associated to similar colors by the software R and this helps in the visual inspection of the 3 clusters. Remember that variable **y** is obtained as linear combination of the covariates **z** and **x** (random and fixed effects respectively). This means that the linear correlation coefficient between **y** and **z** differs across clusters, while the one between **y** and **x** is constant. Therefore, we apply the algorithm allowing both the intercept and one covariate (**z**) to be random[3]. The non-parametric two-level model takes the form:

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

where $i = 1, \ldots, 9$ and $l = 1, \ldots, M$ where M is unknown a priori to the algorithm.

Starting from nine distinct groups, the NPEM algorithm identifies three clusters (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 2.

[3]The algorithm is run considering the following choice of parameters: $D = 0.5$, $\tilde{w} = 0.05$, `it`=30, `it1`=20, `itmax` = 20 and `toolF`=`tollR`= 10^{-4}.

	$\hat{c}_0$	$\hat{c}_1$	$\hat{\beta}$	$\hat{w}$
l=1	5.057	9.942	2.957	1/3
l=2	1.983	4.795	2.957	1/3
l=3	0.261	1.747	2.957	1/3

Table 2: Coefficients of Eq. (19) estimated by the NPEM algorithm. Each row corresponds to a cluster l. The intercept and the coefficient of z differ across groups (c_0 and c_1 respectively), while the coefficient of x (β) is fixed. $\hat{w}$ represents the weight assigned to each cluster.

The estimates of the parameters of random and fixed effects are relatively precise: $(\hat{\mathbf{c}}_0, \hat{\mathbf{c}}_1)$ are very close to $(\mathbf{c}_0, \mathbf{c}_1)$ that we use to generate the data, as well as $\hat{\beta}$ and β ($|\hat{c}_{pl} - c_{pl}| < 0.3$ for $l = 1, 2, 3$ and $p = 1, 2$ and $|\hat{\beta} - \beta| < 0.1$) . Moreover, masses' volumes are proportional to the percentage of data that belongs to each mass. In this case, the algorithm correctly assigns the nine groups to the three clusters, so that, the three volumes are the same since each mass contains $\frac{1}{3}$ of the total number of observations ($\hat{w}_l = 1/3, l = 1, \ldots, 3$). Data with the three identified regression planes are shown in Figure 2.

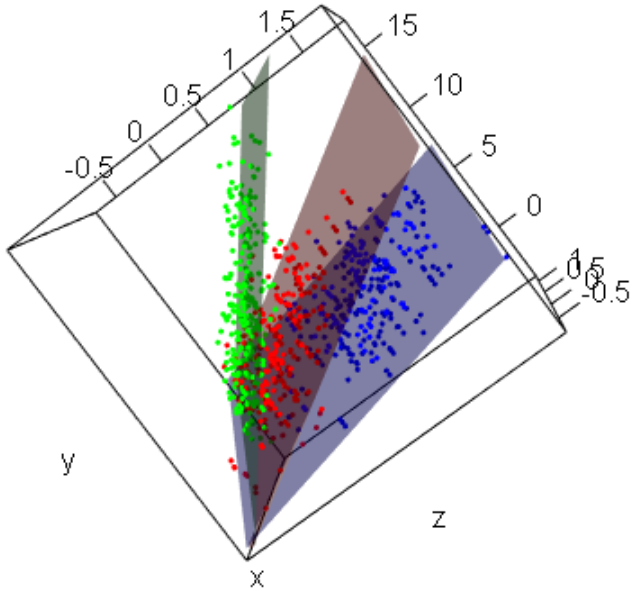

Figure 2: Result of the NPEM algorithm applied to the simulated data according to Equations (18) and (17). Colors represent the three clusters that the algorithm identifies and planes are the estimated linear regression planes within each cluster. Each group is painted with the color of the cluster to which it belongs.

In Figure 2, observations that belong to the same cluster are associated to the same color and, in

this simulation, the algorithm associates each observation to the correct cluster. The three identified regression planes are able to fit the three distinct clouds of data in the most precise way possible. We can conclude that, in this simulation study, the NPEM algorithm is able to identify the latent structure that elapses within the nine groups of data. In particular, it can identify which is the effective number of clusters in which the data are nested and it can characterize each one of these clusters by means of the estimates of the associated parameters.

3 Case study: application of NPEM algorithm to education INVALSI data

In this section, we describe the INVALSI dataset (Section 3.1) and we apply the NPEM algorithm to these data, in order to identify clusters of Italian schools (Section 3.2).

3.1 The INVALSI dataset

INVALSI is an Institute that tests Italian students at different grades and at different years. The data that we analyze in this paper are taken from the INVALSI survey of 2013/2014. Among others, the survey provides several information both at student and at school level. Students, in addition to solve tests in different school subjects, have to fill out a questionnaire about theirselves, their family situations and their habits. Moreover, also school principals have to fill out a questionnaire about himself/herself, his/her school practices and management, school body composition and school size, school structures, infrastructures and school climate. The dataset collects information about $8,946$ students nested within 586 schools. The aim of applying the NPEM algorithm to INVALSI data is that we are interested in exploring the different relations between students 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:

- *MATH8*: student mathematics test score at grade 8 (students attending the last year of junior secondary school in the year 2013/2014);
- *MATH6*: student mathematics test score at grade 6 (students attending the first year of junior secondary school in the year 2011/2012);
- *ESCS*: student socio-economical index.

Student test scores range between 0 and 100, while the ESCS is an indicator built by INVALSI as a continuous variable with mean = 0 and variance = 1. This indicator considers (i) parents' occupation and educational qualifications, and (ii) whether the student owns certain items at home (for instance, the number of books). In general, pupils with an ESCS greater than or equal to 2 are socially and culturally highly advantaged. Figure 3 and Table 3 show variables distributions and descriptive statistics respectively.

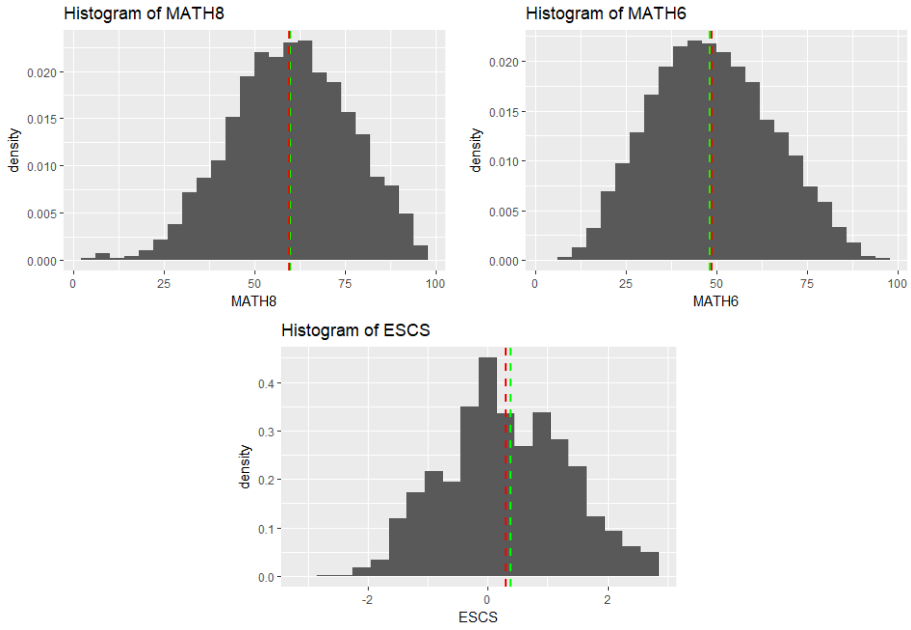

Figure 3: Histograms of students' INVALSI test scores at grade 8, at grade 6 and socio-economical index (ESCS). Red lines refer to the means, green ones to the medians.

	Mean	sd	Median	IQR
$MATH8$	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

Table 3: Descriptive statistics of student level variables employed in the analysis.

Moreover, we have information about the macro-area of localization of schools. About 59% of schools is in Northern Italy, 18% is in Central Italy and 23% is in Southern Italy. Geographical information is a very relevant aspect since many studies in Italy confirm that there are significant discrepancies between students and schools performances across the three geographical macro-areas, [1], [2], [13] and [14].

Since, in a second stage of the analysis, we will look for a characterization of the identified school clusters, Table 4 reports the school level variables that we are interested in, with their descriptive statistics. In particular, variables concern two areas of schools. The former is the *school body composition*: school mean socio-economical index, percentage of females, immigrants, late/early-enrolled students[4], school size and the dummy for private/public school. The latter is related to the *school principal's features*: gender, age, education, years of experience and school practices.

[4]Late/early-enrolled students are those students who started the school grade later or earlier respect to their peers.

Variable Name	Mean	sd	Median	IQR
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 percent	1.21	4.13	0.00	0.00
Late-enrolled student percent	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=1)	70.01%	–	–	–
Age	55.13	7.49	56.00	11.00
Master after degree(yes=1)	22%	–	–	–
Scientific education(yes=1)	14.62%	–	–	–
Year of experience	9.23	7.79	7.00	10.00
Year of experience in the actual school	5.08	5.18	3.00	5.00
Experience in an other district	25.37%	–	–	–
Experience with INVALSI	51.34%	–	–	–
Satisfaction of principal about his/her autonomy [1,10]	7.24	1.81	7.55	1.58

Table 4: School level variables of the database used in the analysis, with their descriptive statistics.

3.2 NPEM algorithm applied to INVALSI data

The aim of this subsection is to apply the EM algorithm for non-parametric mixed-effects models to INVALSI database of 2013/2014 as a tool for clustering Italian schools standing on their student attainments. The correlation between previous student scores (grade 6) and current student scores (grade 8) changes across schools, in the sense that the effects (or values-added) 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 two years before (grade 6) combined with the effect of having attended a particular school for two 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, standing on the results, we group schools into different clusters. In this perspective, the NPEM algorithm works as an in-built classifier, since it performs the grouping of schools into clusters, without knowing a priori the number of clusters.

Standing on previous literature, it is reasonable to think that there is a linear correlation between student scores at grade 6 and at grade 8, [1], [13] and [14]. We therefore consider a non-parametric two-level model (where students represent the first level and schools the second one), with student test scores at grade 6 and student socio-economical index as randon and fixed effects respectively, allowing

both the intercept and the coefficient of student test scores at grade 6 to be random/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 total number of students and $\sum_{i=1}^{N} n_i = J$, the model takes the following form:

$$\begin{aligned} \mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{1}b_{0i} + \mathbf{z}_i b_{1i} + \boldsymbol{\epsilon}_i \qquad & i = 1, \ldots, N \\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{1}_{n_i}) \qquad & ind. \end{aligned} \tag{20}$$

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, while 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-economical index (ESCS) of the n_i students within the i-th school. The choice of considering ESCS as fixed effect and MATH6 as random one is due to the fact that we are interested in exploring how the correlation between MATH6 and MATH8, seen as the reflex of schools ability in training students to achieve certain results, given their students starting potential, varies among schools.

In order to have robust estimates, we select, from the dataset presented in Section 3.1, only the schools that have at least ten students. The resulting dataset consists of 6,188 students nested within 363 schools.

The NPEM algorithm is applied, considering $\tilde{w} = 0.015$, $D = 0.5$, `it`=30, `itmax`=`it1`=20 and `tollR`=`tollF`=10^{-4}. Given these parameters, the algorithm identifies $M = 5$ distinct clusters, whose estimates of parameters are shown in Table 5.

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

Table 5: ML estimates of coefficients of model (20) obtained applying the NPEM algorithm to a selection of INVALSI data of 2013/2014.

The coefficient β in Table 5 is the coefficient related to ESCS (fixed effect). Its positive value (1.417) suggests that, on average, students with high socio-economical index are associated to high performances, in line with previous literature [21]. The estimated $\hat{w}_l$, for $l = 1, \ldots, M$, express the percentage of Italian schools belonging to each cluster l, for $l = 1, \ldots, M$. We identify two main clusters (Cluster 2 and Cluster 3 in Table 5), that contain about the 77% of the total population, while the remaining 23% is distributed across the three other clusters. Regarding the analysis of the coefficients of random effects, Figure 4 helps us in their visualization.

Looking at the figure, it is immediately evident that there is a quite anomalous cluster, identified by lilac color, characterized by a very low slope (Cluster 5 in Table 5). From an interpretative point of view, this cluster contains the "worse" set of Italian schools. Indeed, it is characterized by both low intercept and slope and this means that students in these 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 that had good results at grade 6, after attending two years in a secondary school belonging to Cluster 5, have on average low

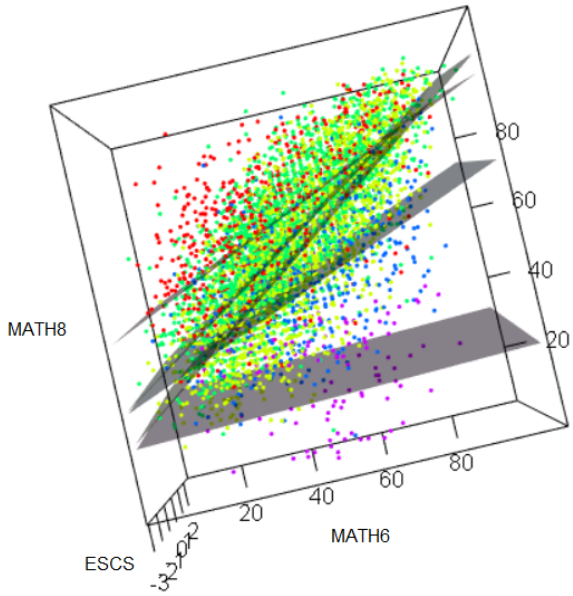

Figure 4: Plot of INVALSI data with the five regression planes identified by the NPEM algorithm , for model (3). Parameters are shown in Table 5. Colors represent the five clusters.

performances, similar to the ones of those students that performed worse than them two years before. On the other side, the best scenario is represented by the cluster on the top of the figure, identified by red color (Cluster 1 in Table 5), that is characterized by a very high intercept (46.028) and a still high slope (0.454). These values suggest that even students that had very low scores at grade 6, obtain high scores at grade 8 with respect to their counterparts attending schools belonging to other clusters. Moreover, the value of the slope suggests that, even if students had on average an improvement on their performances, there is still heterogeneity across students that performed differently two years before, in the sense that best students continue to perform the best with respect to the average.

Thanks to the multilevel structure, we can also compute the Percentage of Variability explained by Random Effects (PVRE), that, in our case, is the percentage of variability in student test scores explained at school level:

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

Given the two-level non-parametric model:

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

the variance of random effects is given by

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

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

In order to provide an index for the goodness of fit of the model, we provide a leave-one-out cross-validation, we compute the Mean Square Error (MSE) and we compare it with the ones obtained considering (i) the same model but with all the parameters as fixed effects and (ii) the parametric mixed-effects models with the same choice of random and fixed effects. Table 6 reports the three MSE computed on the student test scores.

	Parametric FE model	Parametric RE model	NPEM random intercept/slope
MSE	155.91	111.55	118.69

Table 6: Mean Square Error (MSE) computed in three models: (i) parametric fixed-effects model (Parametric FE model); (ii) parametric mixed-effects models with both intercept and covariate as random effects (Parametric RE model); (iii) Non-parametric mixed-effects model with both intercept and slope as random effects (NPEM random intercept/slope).

The MSE obtained with the fixed-effects model is the highest one (155.91) and it departs from the ones obtained by both the parametric and non-parametric mixed-effects models (111.55 and 118.69 respectively). Standing on 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 non-parametric mixed-effects model produces a slightly bigger MSE, but it extrapolates a new kind of information from the data. Indeed, while the parametric approach is able to estimate the parameters of a model, that is based on an already known structure of the data, the non-parametric approach makes a further step, since it is able to identify a new structure within the data, that is the existence of a new, latent, level of grouping. Moreover, the relatively small difference between the MSEs of the two approaches suggests that the clusters structure identified by the NPEM algorithm catches almost all the heterogeneity across the effects of Italian schools.

The further consequence of the identification of a latent structure within the data is that clusters likely derive from some unknown characteristics of schools, that lead to these differences. In a general perspective, the interpretation a posteriori of clusters of data is important per se, especially when speaking about Big Data, where the identification of patterns within a big 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 clusters.

3.2.1 Association between school characteristics and school clusters

Applying the NPEM algorithm to INVALSI data, we discover a structure of clusters that clearly reflects heterogeneities among the effects of Italian schools. In particular, we identify five different clusters, that emerge from five different behaviors of schools in affecting the evolution of their student achievements. We are interested in exploring a posteriori these clusters, in order to investigate whether there are school characteristics that are associated to them. To this end, we analyze the association

of each school level variable presented in Table 4 to the five school clusters.

The first interesting aspect regards geographical differences. Figure 5 reports the proportion of schools belonging to the five clusters, in the three geographical Italian macro-areas: Northern, Central and Southern Italy.

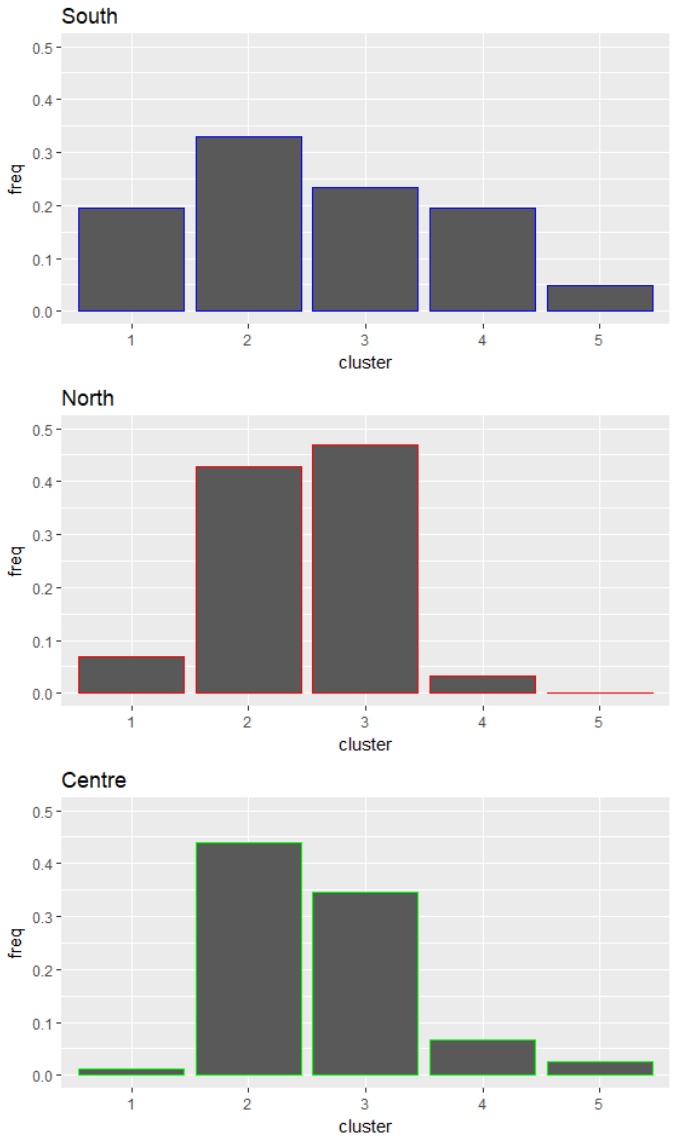

Figure 5: Proportion of schools belonging to the five clusters, within the three geographical Italian macro-areas: North, Centre and South of Italy.

Comparing Northern and Southern Italy, we can notice that the distribution of schools among

clusters is different. In Northern Italy, we do not have any school belonging to Cluster 5 and we have very few schools belonging to Clusters 1 and 4: almost all schools belong to Clusters 2 and 3. In Southern Italy, the distribution of schools among clusters is more homogeneous and it is possible to count a good quantity of schools belonging to each cluster.

Another variable that results to be associated to the distribution of schools among the five clusters is the percentage of immigrant students in school (Immigrant percentage). Figure 6 reports the boxplots of school immigrants percentage within the five distinct clusters.

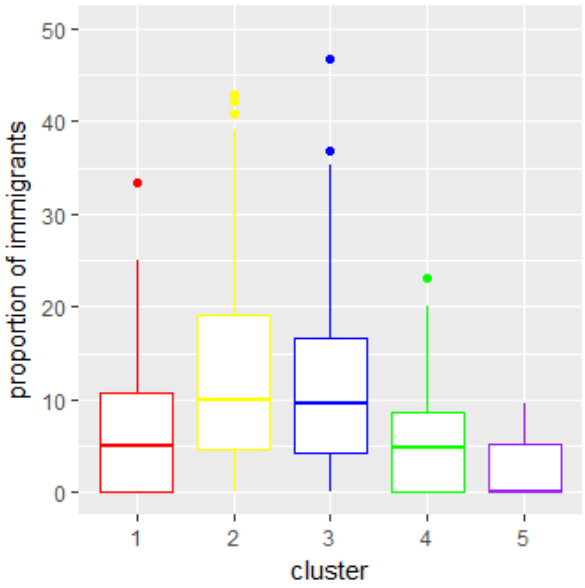

Figure 6: Boxplots of school immigrants percentage within the five clusters. P-value of Kruskall-Wallis test < 0.001.

The two mode clusters (Clusters 2 and 3) contain the higher percentage of immigrant students, while all the other clusters have lower ones. In particular, Cluster 5, that contains schools with the "worse" effect on their student achievements, is characterized by schools with a very low percentage of immigrants. This is in line with previous studies [14], where it emerges that most of immigrant students in Italy are in the North and very few are in the South: being Cluster 5 composed by only schools in Central and Southern Italy, it is reasonable to expect it to have schools with a low percentage of immigrants.

Regarding the entire set of school level variables at our disposal, these two are the only ones that result to be associated to the partition of schools within the estimated five clusters, so that, there is no evidence of significant associations between the other observed school level variables and the clusters. This result does not imply that there is no explanation for the presence of clusters of schools, but, most likely, these clusters derive from other dynamics, that we are not able to observe or to measure.

4 Conclusions

This paper proposes an EM algorithm for non-parametric mixed-effects models (NPEM algorithm), shows a simulation study and applies the NPEM algorithm to INVALSI data of 2013/2014 as a tool for clustering Italian schools. The NPEM algorithm places itself in the literature branch concerning the algorithms proposed in [3] and [4]. In particular, our algorithm is inspired by the one proposed in [4] but it introduces the major improvement, among the others, that the covariates are group specific, meaning that they can vary both in number of observations and range of assumed values across groups. Moreover, with respect to the algorithm proposed in [3], the advantage of NPEM algorithm is that it does not need to fix a priori the number of discrete masses (clusters), but, standing on certain parameters, the algorithm itself identifies the number of discrete support points. This aspect has a great value in the applications where the number of clusters 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 hera of Big Data, where there is the need of identifying latent structures within big and complex databases.

The application to INVALSI data allows us to identify five school clusters that represent different *school effects* on their student achievements, seen as the ability of junior secondary schools in training students to obtain certain skills at the end of the three years, given their skills at the beginning of the school, adjusting for their socio-economical index (ESCS). In the INVALSI framework, schools are associated to *positive or negative value-added*, standing on the final performances of their students and given their students initial skills. Among these five clusters, the presence of a cluster containing schools with a negative value-added is immediately evident. This cluster contains schools that have students which tend to underperform, with respect to their performance two years before, since they have on average very low scores, even if two years before, when they started to attend these schools, they obtained higher scores. Regarding positives value-added, we interpret the cluster with the highest intercept and positive slope (Cluster 1) as the best one, in terms of school effect, since it contains schools able to train students to obtain high performances, even if they had low performances at the beginning of the school. It is worth to say that, from a policy perspective, the definition of the *best school effect* is currently debated. 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, on the other hand, a different point of view emphasizes the advantages of having heterogeneity within the school. In this perspective, the role of the school is to continuously increase the student goals in order to stress the pupils to perform even better, using competition and variation to motivate them.

After the identification of school clusters, the paper focuses on an other actual and interesting topic, that is their interpretation a posteriori. In particular, we explore the associations between school clusters and school level characteristics, showing that only geographical areas and percentage of immigrants result to be significantly associated. This evidence suggests that the school level variables at our disposal do not explain the differences in schools value-added. Standing on the fact that the school clusters are clearly different in their effect on student attainments, the lack of a stratification of school level variables across clusters might means 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 are not able to measure, that might explain the different effects of schools on their students.

In a future perspective, our aim is to deepen the analysis on the characterization of the estimated school clusters, considering other information about the school environment, that we have not been able to measure until now. Moreover, from a methodological point of view, our scope is to develop the

multivariate version of the EM algorithm for non-parametric mixed-effects models, in order to consider two (or more) response variables and to relax the linearity assumptions, considering also the case of other functional forms. In the framework of INVALSI, since the dataset contains both the student scores in reading and 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 standing on both their effects on reading and mathematics student attainments, analyzing the interactions between these two fields.

References

[1] T. Agasisti, F. Ieva, and A. M. Paganoni. Heterogeneity, school-effects and the north/south achievement gap in italian secondary education: evidence from a three-level mixed model. *Statistical Methods & Applications*, 26(1):157–180, 2017.

[2] T. Agasisti and G. Vittadini. Regional economic disparities as determinants of student's achievement in italy. *Research in Applied Economics*, 4(2):33, 2012.

[3] M. Aitkin. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and computing*, 6(3):251–262, 1996.

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

[5] R. D. Bock. *Multilevel analysis of educational data*. Elsevier, 2014.

[6] A. S. Bryk and S. W. Raudenbush. Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, 97(1):65–108, 1988.

[7] P. Clarke, C. Crawford, F. Steele, and A. Vignoles. Revisiting fixed-and random-effects models: some considerations for policy-relevant education research. *Education Economics*, 23(3):259–277, 2015.

[8] J. S. Coleman, E. Campbell, C. Hobson, J. McPartland, A. Mood, F. Weinfeld, and R. York. The coleman report. *Equality of Educational Opportunity*, 1966.

[9] E. A. Hanushek, S. G. Rivkin, and L. L. Taylor. Aggregation and the estimated effects of school resources. Technical report, National bureau of economic research, 1996.

[10] G. Johnes, C. Masci, and T. Agasisti. Student and school performance in the oecd: a machine learning approach. *MOX report*, (26/2017).

[11] B. G. Lindsay et al. The geometry of mixture likelihoods: a general theory. *The annals of statistics*, 11(1):86–94, 1983.

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

[13] C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni. Does class matter more than school? evidence from a multilevel statistical analysis on italian junior secondary school students. *Socio-Economic Planning Sciences*, 54:47–57, 2016.

[14] C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni. Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. *Journal of Applied Statistics*, 44(7):1296–1317, 2017.

[15] C. Nicoletti and B. Rabe. The effect of school spending on student achievement: addressing biases in value-added models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2017.

[16] J. C. Pinheiro and D. M. Bates. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56, 2000.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[18] S. Raudenbush and A. S. Bryk. A hierarchical model for studying school effects. *Sociology of education*, pages 1–17, 1986.

[19] C. Sani and L. Grilli. Differential variability of test scores among schools: A multilevel analysis of the fifth-grade invalsi test using heteroscedastic random effects. *Journal of applied quantitative methods*, 6(4):88–99, 2011.

[20] C. S. Sarrico, M. J. Rosa, and M. J. Manatos. School performance management practices and school achievement. *International Journal of Productivity and Performance Management*, 61(3):272–289, 2012.

[21] S. R. Sirin. Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, 75(3):417–453, 2005.

[22] J. Vanthienen and K. De Witte. *Data Analytics Applications in Education*. Taylor and Francis, 2017.

MOX Technical Reports, last issues

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

62/2017 Barbarotta, L.; Rossi, S.; Dede', L.; Quarteroni, A.
A Transmurally Heterogeneous Orthotropic Activation Model for Ventricular Contraction and its Numerical Validation

61/2017 Vadacca, L.; Colciago, C. M.; Micheletti, S.; Scotti, A.
Three-dimensional fault representation by interface and solid elements: effects of the anisotropy of the fault zone permeability on the timing of triggered earthquakes

60/2017 Bonaldi, F.; Di Pietro, D. A.; Geymonat, G.; Krasucki, F.
A Hybrid High-Order method for Kirchhoff-Love plate bending problems

59/2017 Grujic, O.; Menafoglio, A.; Guang, Y.; Caers, J.
Cokriging for multivariate Hilbert space valued random fields. Application to multifidelity computer code emulation

58/2017 Landajuela, M.; Vergara, C.; Gerbi, A.; Dede', L.; Formaggia, L.; Quarteroni, A.
Numerical approximation of the electromechanical coupling in the left ventricle with inclusion of the Purkinje network

57/2017 Ballarin, F.; D'Amario, A.; Perotto, S.; Rozza, G.
A POD-Selective Inverse Distance Weighting method for fast parametrized shape morphing

56/2017 Alberti, G. S.; Santacesaria, M.
Infinite dimensional compressed sensing from anisotropic measurements

55/2017 Agosti, A.; Cattaneo, C.; Giverso, C.; Ambrosi, D.; Ciarletta, P.
A computational platform for the personalized clinical treatment of glioblastoma multiforme

54/2017 Dede', L; Quarteroni, A.
Isogeometric Analysis of a Phase Field Model for Darcy Flows with Discontinuous Data

53/2017 Bertagna, L.; Deparis, S.;Formaggia, L.; Forti, D.; Veneziani A.
The LifeV library: engineering mathematics beyond the proof of concept