Unsupervised clustering of Italian schools via non-parametric multilevel models

Classificazione non supervisionata delle scuole italiane per mezzo di modelli a effetti misti non parametrici

Chiara Masci, Francesca Ieva and Anna Maria Paganoni

Abstract This work proposes an EM algorithm for the estimation of non-parametric mixed-effects models (NPEM algorithm) and shows its application to the National Institute for the Educational Evaluation of Instruction and Training (INVALSI) dataset of 2013/2014, as a tool for unsupervised clustering of Italian schools. Among the main novelties, the NPEM algorithm, when applied to hierarchical data, allows the covariates to be group specific and assumes the random effects to be distributed according to a discrete distribution 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. In the application to INVALSI data, the NPEM algorithm enables the identification of latent groups of schools that differ in their effects on student achievements.

Abstract *Questo lavoro propone un algoritmo EM per la stima di modelli a effetti misti non parametrici (algoritmo NPEM) e mostra la sua applicazione ai dati dell'Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione (INVALSI) 2013/2014, con l'obiettivo di fare classificazione non supervisionata delle scuole italiane. Tra i principali vantaggi, l'algoritmo NPEM, applicato a dati gerarchici, permette alle covariate di essere specifiche del gruppo e assume che gli effetti casuali seguano una distribuzione discreta, con un numero di masse non noto a priori. Questa assunzione induce un clustering automatico del fattore di raggruppamento al piú alto livello della gerarchia. Nell'applicazione ai dati INVALSI, l'algoritmo NPEM permette l'idetificazione di gruppi latenti di scuole, che differiscono nel loro effetto sul rendimento scolastico degli studenti.*

Chiara Masci
Politecnico di Milano, via Bonardi 9, 20133 Milan e-mail: chiara.masci@polimi.it

Francesca Ieva
Politecnico di Milano, via Bonardi 9, 20133 Milan e-mail: francesca.ieva@polimi.it

Anna Maria Paganoni
Politecnico di Milano, via Bonardi 9, 20133 Milan e-mail: anna.paganoni@polimi.it

Key words: EM algorithm, Non-parametric mixed-effects models, student achievements.

1 Introduction

Administrative educational databases are often characterized by a hierarchical structure, in which students are nested within classes, that are in turn nested within schools. Given this, mixed-effects models are increasingly used in several educational applications. Mixed-effects models include parameters associated with the entire population (fixed effects) and subject/group specific parameters (random effects). They provide both estimates for the entire population's model and for each group's one, where the random effects represent a deviation from the common dynamics of the population. In this work, we develop random effects models, for applying them to educational data, whose random effects have a different meaning: they describe the common dynamics of different clusters of subjects/groups. Indeed, the mixed-effects models that we develop provide estimates for each cluster specific model and they may be considered as an unsupervised clustering tools for hierarchical data. The difference with respect to classical parametric mixed-effects models is that the random effects, instead of being Normal distributed, follow a discrete distribution that we call P^* [16]. Most of the mixed-effects models used in the educational field are parametric linear multilevel models [6], but parametric assumptions sometimes result to be too restrictive to describe very heterogeneous populations. Moreover, when the number of measurements for group is small, predictions for random effects are strongly influenced by the parametric assumptions. For these reasons, we opt for a nonparametric (NP) framework, which allows P^* to live in an infinite dimensional space and that also provides, in a natural way, a classification tool. Hierarchical models have been already applied to educational data in the Italian literature: [1], [2], [11] and [17] apply multilevel linear models in order to disentangle the portion of variability in students' scores given to different levels such as the family, the class or the school. Differently, our algorithm aims at identifying clusters of schools that perform in similar ways and, in a second step, at characterizing these clusters in terms of similarities within/between groups [13]. To the best of our knowledge, this is one of the first times that this kind of algorithm has been applied in the educational context [7]. Our method is strictly related to the branch of literature about growth mixture models (GMM) [15], latent class analysis (LCA) [14] and finite mixture models [18], which also aim at the identification of latent subpopulations, but with the main difference that all these models need to fix a priori the number of latent subpopulations. The choice of the number of latent classes (mass points) is not trivial when the sample is very big or the knowledge about possible different trends across the individuals (groups) is limited. For this reason, our approach brings a significant value-added with respect to the existing literature. In particular, our algorithm is inspired by both the one proposed in [3] and [4] and the one proposed in [5], but with some substantial changes. Contrarily

to the algorithm described in [3] and [4], we do not need to fix the number of groups a priori but the algorithm identifies it by itself, standing on given tolerance values. While referring to the algorithm in [5], we adjust it in order to consider the linear case, to allow the covariates to be group-specific and to compute the optimization of the Maximization step in closed-form. We apply this algorithm to INVALSI data of year 2013/2014, in which we consider students nested within schools. Each group is identified by a school and the aim is to cluster schools standing on their different effects on their student performance trends. In this way, it is possible to identify clusters of schools that perform in different ways, trying to find out which are the determinants of different school effects.

2 The Dataset

The INVALSI database [8] contains information about more than 6,500 Italian students attending the third year of junior secondary school in the year 2013/2014, nested within about 500 schools. At pupil's level, we have reading and mathematics INVALSI test scores at grade 8 (RS and MS) and also, reading and mathematics INVALSI test scores at grade 6, two years before, of the same students. It is well known from the literature that education is a cumulative process, where achievement in the period t exerts an effect on results of the period t + 1. These variables take values between 0 and 100. Moreover, the following information is available: gender, immigrant status, if the student is early/late-enrolled, information about the family's background and socioeconomic status of the student (ESCS). At school's level, we have variables about three different areas: (i) the school-body composition (school-average characteristics of students, such as the proportion of immigrants, early and late-enrolled students, etc); (ii) school principal's characteristics; (iii) managerial practices of the school. Two dummies are also included to distinguish (i) private schools from public ones, and (ii) "Istituti Comprensivi", which are schools that include both primary and lower-secondary schools in the same building/structure. This latest variable is relevant to understand if the "continuity" of the same educational environment affects (positively or negatively) students results. Some variables about size (number of students per class, average size of classes, number of students of the school) are also included to take size effects into account. Lastly, regarding geographical location, we include two dummies for schools located in Central and Southern Italy and the district in which the school is located; some previous literature, indeed, pointed at demonstrating that students attending the schools located in Northern Italy tend to have higher achievement scores than their counterparts in other regions, all else equal. As we have the anonymous student ID, we have also the encrypted school IDs that allow us to identify and distinguish schools.

3 Methodology

We consider the case of a non-parametric two-level model with one covariate with fixed effect, one covariate with random slope and one random intercept. The model takes the following form:

$$\begin{aligned} &\mathbf{y}_i = \beta \mathbf{x}_i + c_{0l} + c_{1l}\mathbf{z}_i + \varepsilon_i \quad i = 1,\dots,N, l = 1,\dots,M \\ &\varepsilon_i \overset{ind}{\sim} N(\mathbf{0}, \sigma^2 \mathbf{1}_n) \end{aligned} \tag{1}$$

where, in our application, N is the total number of schools; $\mathbf{y}_i$ is the n_i-dimensional vector of student achievements at grade 8 in school i; $\mathbf{x}_i$ is the n_i-dimensional vector of ESCS of students in school i; $\mathbf{z}_i$ is the n_i-dimensional vector of the same students achievements at grade 6 (two years before) in school i. We use these three variables at student level and we make this choice of random and fixed effects because we are interested in modeling the association between student achievements at grade 6 and 8, across different schools, adjusting the model for the effect of the ESCS, that, standing on the Italian literature, [2], [11], [12], results to be one of the most influential variable, with an homogeneous effect in the whole country. $\mathbf{c} \in R^2$ is the vector containing the coefficients of random effects. $\mathbf{c}$ follows a discrete distribution P^* with M support points, where M is not known a priori. P^* can then be interpreted as the mixing distribution that generates the density of the stochastic model in (1). The ML estimator $\hat{P}^*$ of P^* can be obtained following the theory of mixture likelihoods in [9] and [10], 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.The ML estimator of the random effects distribution can be expressed as a set of points $(c_1,\dots,c_M)$, where $M \leq N$ and $c_l \in R^2$ 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 develop an algorithm for the joint estimation of σ^2, β, $(c_1,\dots,c_M)$ and $(w_1,\dots,w_M)$, that is performed through the maximization of the likelihood, mixture by the discrete distribution of the random effects,

$$L(\beta,\sigma^2,c_l,w_l|y) = \sum_{l=1}^{M} \frac{w_l}{(2\pi\sigma^2)^{\sum_{i=1}^{N} n_i}} exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} (y_{ij} - \beta x_{ij} - c_{0l} - c_{1l} z_{ij})^2 \Big\} \tag{2}$$

with respect to σ^2, β and (c_l, w_l), for $l = 1,\dots,M$. Each school i, for $i = 1,\dots,N$ is therefore assigned to a cluster l, for $l = 1,\dots,M$. The 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. Moreover, given N starting support points, during the iterations of the EM algorithm, we reduce the support of the discrete distribution standing on both two criteria: the former is that we fix a threshold D and

if two points are closer than D they collapse to a unique point; the latter is that we remove points with very low weight ($w_l \leq \tilde{w}$, being $\tilde{w}$ a given threshold on weights) and that are not associated to any school. When two points $\mathbf{c}_l$ and $\mathbf{c}_k$ collapse to a unique point, because their Euclidean distance is smaller than D, we obtain a new mass point $\mathbf{c}_{l,k} = \frac{\mathbf{c}_l + \mathbf{c}_k}{2}$ with weight $w_{l,k} = w_l + w_k$. The thresholds D and $\tilde{w}$ are two complexity parameters that affect the estimation of the nonparametric distribution: the higher is D, the lower is the number of clusters. The choice of the values for D and $\tilde{w}$ depends on how much we want to be sensitive to the differences among clusters (D) and which is the minimum number of groups (schools) that we allow within each clusters ($\tilde{w}$). Anyway, different results obtained using different set of tuning parameters can be compared in terms of AIC or BIC in order to choose the best one. Notice that the number of support points M is computed by the algorithm as well and we do not have to fix it a priori. Since we do not have to specify a priori the number of support points, the NP mixed-effects model could be interpreted as an unsupervised clustering tool for longitudinal data.

4 Results

The algorithm cluster the Italian schools within 5 clusters, whose estimated parameters are shown in Table 1.

	$\hat{\beta}$	$\hat{c}_0$	$\hat{c}_1$	$\hat{w}$
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 1 Coefficients of Eq. (1) 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.

Each cluster is characterized by an intercept, a slope of the grade 6 test score variable and by the fixed coefficient of the ESCS. We identify two main clusters (Cluster 2 and Cluster 3 in Table 1), that contain about the 77% of the total population of schools, while the remaining 23% is distributed across the other three clusters. From an interpretative point of view, with respect to Cluster 2 and Cluster 3 that form the reference cluster, while Cluster 5 contains the "worse" set of Italian schools. Indeed, it is characterized by both low intercept and slope and this means that there is a kind of equality in student achievements, but with on average very low scores at grade 8, even if the results at grade 6 were on average higher. In a second step, we apply a multinomial logit model at school level, by treating the five clusters as the categorical outcome variable and all the school level characteristics as covariates, with the aim of characterizing the identified clusters by means of school

level variables. It emerges that the dummy for private/public school, the percentage of disadvantaged students and the geographical area are associated to heterogeneity across groups.

References

1. Agasisti, T., Vittadini, G.: Regional economic disparities as determinants of student's achievement in Italy. Research in Applied Economics, 4(2), 33 (2012).
2. Agasisti, T., Ieva, F., Paganoni, A.M.: 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).
3. Aitkin M.: A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and computing, 6(3), 251-262 (1996).
4. Aitkin M.: A general maximum likelihood analysis of variance components in generalized linear models. Biometrics, 55(1), 117-128 (1999).
5. Azzimonti, L., Ieva, F., Paganoni, A.M.: Nonlinear nonparametric mixed-effects models for unsupervised classification. Computational Statistics, 28(4), 1549-1570 (2013).
6. Fox, J.: Linear mixed models. Appendix to An R and S-PLUS Companion to Applied Regression (2002).
7. Gnaldi, M., Bacci, S., Bartolucci, F.: A multilevel finite mixture item response model to cluster examinees and schools. Advances in Data Analysis and Classification, 10(1), 53-70 (2016).
8. INVALSI website
   http://www.invalsi.it/
9. Lindsay, B.: The geometry of mixture likelihoods: a general theory. The Annals of Statistics, 11(1), 86-94 (1983).
10. Lindsay, B.: The geometry of mixture likelihoods, part II: the exponential family. The Annals of Statistics, 11(3), 783-792 (1983).
11. Masci, C., Ieva, F., Agasisti, T., Paganoni, A.M.: 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).
12. Masci, C., Ieva, F., Agasisti, T., Paganoni, A.M.: Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. Journal of Applied Statistics, 44(7), 1296-1317 (2017).
13. Masci, C., Ieva, Paganoni, A.M.: Non-parametric mixed-effects models for unsupervised classification of Italian schools. MOX-report 63/2017.
14. McCulloch, C. E., Lin, H., Slate, E. H., Turnbull, B. W.: Discovering subpopulation structure with latent class mixed models. Statistics in medicine, 21(3), 417-429 (2002).
15. Muthén, B., Shedden, K.: Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics, 55(2), 463-469 (1999).
16. Skrondal, A., Rabe-Hesketh, S.: Multilevel and related models for longitudinal data. In Handbook of multilevel analysis (pp. 275-299). Springer, New York, NY (2008).
17. Sulis, I., Porcu, M.: Assessing divergences in mathematics and reading achievement in italian primary schools: A pro posal of adjusted indicators of school effectiveness. Social Indicators Research, 122(2), 607-634 (2015).
18. Tutz, G., Oelker, M. R.: Modelling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures. International Statistical Review, 85(2), 204-227 (2017).