Computational Statistics (2021) 36:2337–2377
https://doi.org/10.1007/s00180-021-01107-1

ORIGINAL PAPER

Evaluating class and school effects on the joint student achievements in different subjects: a bivariate semiparametric model with random coefficients

Chiara Masci[1] · Francesca Ieva[1] · Tommaso Agasisti[2] · Anna Maria Paganoni[1]

Received: 5 October 2020 / Accepted: 19 April 2021 / Published online: 16 May 2021

Abstract
This paper proposes an innovative statistical method to measure the impact of the class/school on student achievements in multiple subjects. We propose a semiparametric model for a bivariate response variable with random coefficients, that are assumed to follow a discrete distribution with an unknown number of support points, together with an Expectation-Maximization algorithm—called BSPEM algorithm—to estimate its parameters. In the case study, we apply the BSPEM algorithm to data about Italian middle schools, considering students nested within classes, and we identify subpopulations of classes, standing on their effects on student achievements in reading and mathematics. The proposed model is extremely informative in exploring the correlation between multiple class effects, which are typical of the educational production function. The estimated class effects on reading and mathematics student achievements are then explained in terms of various class and school level characteristics selected by means of a LASSO regression.

Keywords Semiparametric model · Random coefficients · EM algorithm · School and class effects · Student achievements · Teaching practices

1 Introduction and motivation

Student learning is a long and complex process that sees many different factors acting on it. During their careers, students receive inputs from their family, their peers and

✉ Chiara Masci
chiara.masci@polimi.it

[1] MOX - Department of Mathematics, Politecnico di Milano, P.za Leonardo Da Vinci 32, Milano, Italy

[2] DIG - Department of Management, Economics and Industrial Engineering, Politecnico di Milano, via Lambruschini 4/b, Milano, Italy

the school and class they are attending. The educational system is hierarchical, i.e. different levels of grouping are nested within each others: students are nested within classes, that are in turn nested within schools, that are in turn nested within districts and so on so forth. Each one of these levels has a specific role in the student learning process. Measuring how much of the variability in student education is due to each grouping level of the hierarchy is not easy, but, it is essential for evaluating the role of educational institutions (i.e., schools). In particular, there is a broad and rich literature about *school value-added* based on test scores, intended as the difference in test performance of students in a school and the average performance of schools populated by students with a comparable level of prior achievement (and other student characteristics) (Raudenbush and Willms 1995; Schagen and Schagen 2005; Timmermans et al. 2014). School value-added promises to enable fair comparisons of school performance despite schools having markedly different pupil intakes. The logic behind it is indeed to compare schools only on the basis of unexplained variation between (statistically) “like-for-like” pupils. A simple approach is to compare the performance of a particular group of pupils to the performance of other pupils with the same examination score at the earlier point in time. Beyond prior attainment, there are other non-school factors associated with students’ progress, like socioeconomic status, gender or ethnicity. The inclusion of these confounding variables in the measurement of school value-added has been long debated (Meyer 1997; Strand 1997; McCaffrey et al. 2004; Martineau 2006). The most recent literature about this topic (Perry 2016; Leckie and Goldstein 2017; Parsons et al. 2018) supports the development of the so called *contextual value-added*, that takes into account, besides student test scores, also age, gender, ethnicity, socioeconomic status and various other pupil characteristics when measuring the school value-added. The rationale for *contextual value-added* is that ignoring these contextual factors considerably biases the results, attributing successes and failures to schools inappropriately.

Even though the measurement of school value-added continuously receives attention, decades of educational effectiveness confirm that differences between pupils is more within schools than between them (Hanushek 1992; Perry 2016; Rivkin et al. 2005; Rockoff 2004). In this perpective, the concept of *school value-added*, as intended before, can be transfered to the class level, speaking about *class value-added*. Class peers, class climate and, especially, teachers considerably affect the student learning process. Indeed, different types of teaching practices promote different cognitive skills in students (Bietenbeck 2014) and, now that traditional teaching practices co-exist together with more modern teaching methods (work in small groups, emphasize real-life application), their effects can be very heterogeneous. In the last twenty years, the analysis of teaching practices and effetiveness is increasingly receiving attention and recent studies find evidence of an association between the effects on student achievements and different teaching practices, in different school subjects (Goldhaber and Brewer 1997; Wenglinsky 2002; Schwerdt and Wuppermann 2011; De Witte and Van Klaveren 2014). Focusing on the specific way in which teaching is organized is also important because it allows moving from exploring simple correlations between students results and teachers’ characteristics to a more complex and complete scenario.

In the perspective of evaluating school and class value-added, rich linked national data that contain longitudinal observations are extremely useful. In Italy, the National

Institute for the Educational Evaluation of Instruction and Training (INVALSI) tests students at different grades and at different years, both in reading and mathematics, by means of standardized tests in the entire country. Students are tested at grades II and V of primary school, at grade III of junior secondary school and at grade II of upper secondary school. Moreover, INVALSI collects information about students, teachers, classes, schools and school principals, by means of dedicated questionnaires. In so doing, it creates a dataset that contains a rich picture of the personal and educational reality of each student. This dataset allows to compare the performances of students that attend different classes, in different schools, in the various geographical Italian regions, but with the same yardstick.

The INVALSI dataset has been recently studied by economists and statistical scholars interested in analyzing the determinants of student, class and school performances. In Agasisti et al. (2017); Grilli and Rampichini (2009), Masci et al. (2016, 2017), Sani and Grilli (2011), the authors, considering the hierarchical nature of educational data, apply mixed-effects linear models in order to identify which are the student characteristics associated to student performances and to estimate how much of the variability in student performances is due to their grouping in different classes and schools. These are some of the first attempts that aim at separating and estimating the effects of different levels of grouping on Italian student achievement. In Masci et al. (2016, 2017), the authors apply a three-level hierarchical structure in which students are nested within classes that are in turn nested within schools and measure the contribute of each of these levels on students test scores' variability. Results show that, after adjusting for student characteristics, the variability among student achievements explained at class level is much higher that the one explained at school level. By means of parametric mixed-effects linear models, they estimate the school and class effect, interpreted as the value-added that each school or class gives to the performances of its students. A relevant result that the study in Masci et al. (2017) shows is that the correlation between the school effects on reading and mathematics student achievements is positive and statistically significant, while the correlation between the two class effects is null. This important finding suggests that the effect of the school is most of the times coherent on the different school subjects, probably due to certain school characteristics that have similar effects on different subjects (for example, school principal practices, school body composition and school peers). On the other way, the fact that the correlation among class effects in reading and mathematics is null suggests that there is not a unique effect of the class environment on the different school subjects, but the effects of the class on the two school subjects are potentially uncorrelated. One of the most likely interpretation of this result is that a significant part of the class effect is due to something that is not common between the two school subjects, the main candidate for this being the teachers. Being the teachers in mathematics and reading different, their characteristics and their teaching practices might be completely different too, leading to uncorrelated effects on student achievements.

Our paper aims at estimating the *class effect* in the context of within-school heterogeneity. We follow the approach presented in Masci et al. (2019), where the authors apply a multilevel linear model to estimate the school effect, but, instead of following a classical parametric approach, they follow a semi-parametric approach: they develop a semi-parametric mixed-effects (two-level, where students are nested within schools) model able to identify a latent structure among the highest level of the hierarchy (schools). They cluster schools standing on the evolution of their student achievements across years. In this sense, the concept of *school effect*, re-defined from a methodological point of view, reflects the different effects of schools on the evolution of their student achievements at different grades. In particular, they identify subpopulations of schools within which student mathematics test scores trends (measured by the linear relation between INVALSI test scores at different grades) are similar and, in a second step, they characterize *a posteriori* the identified subpopulations of schools by means of school level characteristics.

In this paper, we extend the statistical model presented in Masci et al. (2019) and we propose a study that is innovative from a methodological and an interpretative point of view. We extend the Expectation-Maximization algorithm for semi-parametric models with random coefficients (SPEM algorithm) presented in Masci et al. (2019) to the bivariate case, i.e. to the case of a bivariate response variable (which, in our case, is the test score in reading and mathematics). We are interested in estimating the impact that attending different classes has on student performance trends, i.e. student performance evolution over time, and, in particular, in comparing these effects between reading and mathematics. With *class effect*, we intend the way in which achievements of students have evolved after attending three years of junior secondary school in a specific class (within a given school). The model that we propose is a bivariate two-level linear model where the random coefficients, under semi-parametric assumptions, follow a bivariate discrete distribution with an unknown number of mass points. Each group is assigned to a bivariate subpopulation of groups, that is represented by specific values of the parameters of the bivariate semi-parametric linear model. The distribution of the random coefficients is a bivariate discrete distribution where each dimension is allowed to have a different finite number, unknown a priori, of mass points. This formulation permits to estimate the marginal distribution of the random coefficients related to each one of the two response variables and, moreover, to estimate the joint distribution of random coefficients related to the two response variables, investigating the correlation among them. Read in the context of the educational literature on school value-added, this method has two main advantages: (i) for the first time the *effect* estimated considers not only heterogeneity within schools (i.e. between classes) but also within classes (i.e. between teachers); (ii) besides the random intercept, that is typically the unique random effect considered in the educational literature on school/class value-added, the inclusion of a random slope allows to model the school/class effect in a more sophisticated way (i.e. modelling the heterogeneity in the association between previous and current student test scores across schools/classes).

Multivariate multilevel models have been frequently used in the educational literature to estimate school and class effects (see, among the others, Yang et al. (2002); Masci et al. (2017)). By assuming Gaussian random effects in multilevel models, we can extract a point estimate for each group (school or class), together with its confidence interval. This setting provide a ranking of the groups where all groups have the same weight and their effects can be compared by looking at their estimated random effects and relative confidence intervals. By assuming discrete random effects in a semi-parametric approach, we identify a latent structure of subpopulations in which groups are clustered. This approach provides an alternative to the ranking that presents several advantages (Rights and Sterba 2016). First of all, the semiparametric approach, being more flexible and not assuming *a priori* any parametric distribution, can estimate the real distribution of the random effects. Secondly, in a context in which the number of groups is extremely large, the identification of subpopulations might help in interpreting the results. Sequential groups in the ranking, whose confidence intervals are overlapped, do not statistically differ and considering their heterogeneity might unnecessarily increase the problem complexity and be misleading. Last but not least, the identification of subpopulations can help in the outlier identification: the most populated subpopulations reveal which are the reference trends, while the smaller subpopulations contain those groups whose observations tend to have anomalous behaviors with respect to the majority. In this perspective, we do not create a full ranking of the highest level effects, but instead we generate subpopulations of effects and we attribute each group to a single subpopulation.

The proposed methodology is new to the literature. The semi-parametric mixed-effects linear model in Masci et al. (2019) on which we base our multivariate model enters in the research line about the identification of subpopulations of the Growth Mixture Models (GMM) (Muthén 2004; Muthén and Shedden 1999; Nagin 1999) and of Latent Class Mixture Models (LCMM) (McCulloch et al. 2002; Vermunt and Magidson 2002), but with the novelty that it does not need to fix a priori the number of latent subpopulations to be identified. Moreover, being the existing methods specified in the Structural Equation Modeling (SEM) framework, they are still relatively limited when covariates are group-specific. Numerous extensions and applications of GMM and LCMM has been already realized (Lin 2000; Muthén and Asparouhov 2015), but none of them include the modeling of a multivariate answer variable, where the latent subpopulations structure of groups (higher level of hierarchy) are allowed to differ across the responses, i.e. are response-specific. Our proposed model is the new extension to the bivariate case of a model that is already innovative by itself and particularly useful in the case of education, where the output is typically multivariate.

The main advantages of the multivariate modelling rely on two aspects. First, considering that the multiple response variables come from a single subject, we expect them to be somehow correlated. The multivariate model takes into account this source of correlation when it estimates the model parameters and, therefore, we expect it to be more appropriate than independent univariate models. Second, the multivariate model allows to estimate the joint distribution of the random effects from which we can investigate the correlation among them, which is of our interests. Fitting independent univariate models would lead to separate univariate distributions of random effects and measuring a posteriori their correlation represents only a raw proxy of the real joint distribution (Leckie 2018).

In this specific paper, our data provided by INVALSI refer to a sample of classes, representative at national level - but one per school, so we cannot estimate the class effects within schools. In other words, our model here is applicated with two-levels (students and classes). The model estimates a bivariate effect for each class, i.e. the effect of the class on mathematics student achievement trends and the one on reading student achievement trends. The aim is to identify how many different trends exist in student performances across classes, for both mathematics and reading, i.e. to identify how many and which are the mass points of the discrete distribution of random coefficients (class effects) for both the first and the second response. Moreover, by looking at the joint distribution of these random coefficients, we investigate the correlation between the class effects on reading and mathematics, allowing differences between them (i.e. assuming that teachers' ability and effectiveness can be different between teachers of the same class).

Therefore, the main research questions that we aim to address are:

- Are there differences across the effects of the Italian classes on their students achievement?
- Are the effects of the classes in reading and mathematics achievements correlated?
- Is it possible to identify groups of classes that perform differently from the majority?
- Do the identified groups of classes differ in terms of class level features, for example teachers characteristics, teaching practices and class body composition?

In the year 2016/2017, INVALSI submitted questionnaires to teachers about their personal information, their education, their teaching practices and the environment of the class and school in which they work, creating an informative and new dataset that, until now and in this context, has been poorly explored. We leverage this brand new opportunity by using this additional information to explore the potential determinants of the class/school effects. In this perspective, in order to investigate whether the different student achievement trends across classes are related to these aspects, in a second stage of the analysis, we look for associations between class and teacher level characteristics and the identified subpopulations of class effects, by means of a lasso multinomial logit model. The questionnaire has been realized only in 2016/2017, so our study is cross-sectional by design.

This paper brings important innovations to the literature on assessment of education results for at least two main aspects. First, it proposes a novel statistical method to perform in-built, unsupervised clustering of the higher level of grouping of a bivariate multilevel model, without knowing a priori the number of clusters (so avoiding the typical rigidities when specifying an educational production function). Second, exploring differences and similarities of class effects in mathematics and reading by means of a multivariate model is a great advantage, also when the bivariate class effects are characterized, in a second step, in terms of class features (teacher characteristics and practices).

The paper is organized as follows: in Sect. 2, we present the bivariate semi-parametric two-level linear model. In Sect. 3, we perform a simulation study. In Secion 4, we focus on the case study, (i) presenting the dataset about Italian middle schools, (ii) applying the EM algorithm for bivariate semi-parametric models with random coefficients - BSPEM algorithm - to it and showing its results and (iii) analyzing a posteriori the characteristics of the identified subpopulations of classes. In Sect. 5, we draw policy implications and conclusions.

2 Model and methods: the bivariate semi-parametric linear model with random coefficients

In this section, we present the bivariate semi-parametric linear model with random coefficients[1].

Consider a bivariate two-level linear model, where each bivariate 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:

$$\begin{aligned} \begin{pmatrix}\mathbf{y}_{1,i}\\ \mathbf{y}_{2,i}\end{pmatrix}^T &= \mathbf{X}_i \begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}^T + \mathbf{Z}_i \begin{pmatrix}\boldsymbol{\delta}_{1,i}\\ \boldsymbol{\delta}_{2,i}\end{pmatrix}^T + \begin{pmatrix}\boldsymbol{\epsilon}_{1,i}\\ \boldsymbol{\epsilon}_{2,i}\end{pmatrix}^T \qquad i = 1, \dots, N, \\ \boldsymbol{\epsilon}_i^T &= \begin{pmatrix}\boldsymbol{\epsilon}_{1,i}\\ \boldsymbol{\epsilon}_{2,i}\end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind. \end{aligned} \tag{1}$$

The components of model (1) are the following[2]:

- $\mathbf{Y}_i = \begin{pmatrix} y_{1,1i}, \dots, y_{1,n_i i} \\ y_{2,1i}, \dots, y_{2,n_i i}\end{pmatrix}^T$ is the $(n_i \times 2)$-dimensional matrix of response variable within the i-th second level group[3],
- $\mathbf{X}_i$ is the $(n_i \times (P+1))$-dimensional matrix of covariates relative to fixed coefficients,
- $\mathbf{B} = \begin{pmatrix}\boldsymbol{\beta}_1 & \boldsymbol{\beta}_2\end{pmatrix}$ is the $((P+1) \times 2)$-dimensional matrix of coefficients of $\mathbf{X}$,
- $\mathbf{Z}_i$ is the $(n_i \times (R+1))$-dimensional matrix of covariates relative to random coefficients,
- $\mathbf{1}_i = \begin{pmatrix}\boldsymbol{\delta}_{1,i} & \boldsymbol{\delta}_{2,i}\end{pmatrix}$ is the $((R+1) \times 2)$-dimensional matrix of random coefficients of $\mathbf{Z}_i$,
- $\boldsymbol{\epsilon}_i = \begin{pmatrix}\boldsymbol{\epsilon}_{1,i} & \boldsymbol{\epsilon}_{2,i}\end{pmatrix}$ is the $(n_i \times 2)$-dimensional matrix of errors and $\boldsymbol{\Sigma}$ is its variance/covariance matrix.

Fixed effects are identified by parameters associated to the entire population, while random ones are identified by group-specific parameters. In the perspective of the application to INVALSI data, this model will consider a two-levels hierarchy: students as level 1 and classes as level 2. In the parametric framework of bivariate linear mixed-effects models, the random coefficients $\boldsymbol{\Delta}_i$ are assumed to be distributed according to a Normal distribution with mean vector equal to $\mathbf{0}$ and a variance/covariance matrix that is estimated, together with the other parameters of the model, through methods based on the maximization of the likelihood or the restricted likelihood functions (Pinheiro and Bates 2000). For each response variable, this parametric distribution allows to associate each group i to a different set of coefficients $\boldsymbol{\delta}_{*,i} = (\delta_{*,i1}, \dots, \delta_{*,i(R+1)})$ for the $(R+1)$ covariates of the random effects, extracted from the normal distribution.

[1] Details about the EM algorithm for the estimation of model parameters and the sketch of the BSPEM algorithm can be found in the Appendix A.

[2] In subscript of each variable/parameter, we indicate by the number before the comma whether the variable/parameter is referred to the first or the second response variable (for example, $y_{1,ji}$ and $y_{2,ji}$ are the j-th first and second response variables within (level 2)-group i, respectively).

[3] We consider the case in which the number of observations of the two response variables is the same within each group, but is allowed to be different across the groups.

Following the idea presented in Masci et al. (2019), we relax the parametric assumptions about the coefficients of the random effects and we assume the bivariate coefficients $\mathbf{1}_i = (\boldsymbol{\delta}_{1,i} \quad \boldsymbol{\delta}_{2,i})$ to follow a bivariate discrete distribution S^*, assuming $M \times K$ mass points $(\mathbf{C}_{11}, \ldots, \mathbf{C}_{MK})$, where each $\mathbf{C}_{mk}$ is the $2 \times (R+1)$-dimensional matrix of coefficients of random effects for the bivariate mass point related to the index (m, k), for each $m = 1, \ldots, M$ and $k = 1, \ldots, K$, where both M and K are smaller than N. The total number of mass points, that is $M \times K$, is unknown a priori and it is estimated together with the other parameters of the model. This modelling allows the identification of a bivariate clustering distribution among the N groups, where each group i is associated to a bivariate cluster, standing on the linear relationships between the two response variables and their covariates. In other words, the model identifies a bivariate latent structure among the groups, that also reveals the dependence among the two response variables. Under these assumptions, the semi-parametric bivariate model with random coefficients takes the following form:

$$\begin{gathered}
\begin{pmatrix} \mathbf{y}_{1,i} \\ \mathbf{y}_{2,i} \end{pmatrix}^T = \mathbf{X}_i \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}^T + \mathbf{Z}_i \begin{pmatrix} \mathbf{c}_{1,m} \\ \mathbf{c}_{2,k} \end{pmatrix}^T + \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix}^T \\
i = 1, \ldots, N \qquad m = 1, \ldots, M \qquad k = 1, \ldots, K \\
\boldsymbol{\epsilon}_i^T = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind.
\end{gathered} \tag{2}$$

Without loss of generality, we consider the case of a semi-parametric bivariate two-level linear model, with one random intercept, one random covariate and P fixed covariates[4]. Model (2) reduces to:

$$\begin{gathered}
\begin{pmatrix} \mathbf{y}_{1,i} \\ \mathbf{y}_{2,i} \end{pmatrix}^T = \mathbf{1}_{n_i} \begin{pmatrix} c_{1,1m} \\ c_{2,1k} \end{pmatrix}^T + \sum_{p=1}^{P} \mathbf{x}_{ip} \begin{pmatrix} \boldsymbol{\beta}_{1p} \\ \boldsymbol{\beta}_{2p} \end{pmatrix}^T + \mathbf{z}_i \begin{pmatrix} c_{1,2m} \\ c_{2,2k} \end{pmatrix}^T + \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix}^T \\
i = 1, \ldots, N \qquad m = 1, \ldots, M \qquad k = 1, \ldots, K \\
\boldsymbol{\epsilon}_i^T = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind.
\end{gathered} \tag{3}$$

where $\mathbf{1}_{n_i}$ is the n_i-dimensional vector of 1, M is the total number of mass points for the first response and K is the total number of mass points for the second response and both of them are unknown a priori. Coefficients $\mathbf{C}_{mk}$, for $m = 1, \ldots, M$ and $k = 1, \ldots, K$ are distributed according to a discrete probability measure S^* that belongs to the class of all probability measures on $\mathcal{R}^4$. S^* can then be interpreted as the mixing distribution that generates the density of the stochastic model in (3). The ML estimator $\hat{S}^*$ of S^* can be obtained following the theory of mixture likelihoods in Lindsay (1983a, b), as explained in Masci et al. (2019). In particular, in Lindsay (1983a, b), the authors prove the existence, discreteness and uniqueness of the semiparametric maximum likelihood estimator of a mixing distribution, in the

[4] This choice is driven by the application in the case study shown in Sect. 3. Nonetheless, the BSPEM algorithm allows to consider as random effects both the intercept and one slope, as well as only one of them.

case of exponential family densities. Proofs of the identifiability property can be found in Teicher (1963); Barndorff-Nielsen (1965). The ML estimator of the random coefficients distribution can be expressed as a set of points $(\mathbf{C}_{11}, \ldots, \mathbf{C}_{MK})$ and a set of weights $(w_{11}, \ldots, w_{MK})$, where $\sum_{m=1}^{M} \sum_{k=1}^{K} w_{mk} = 1$ and $w_{mk} \geq 0$, for $m = 1, \ldots, M$ and $k = 1, \ldots, K$. Each group i, for $i = 1, \ldots, N$, is assigned to a bivariate cluster (m, k), standing on the fact that the first response belongs to cluster m and the second one to cluster k. Indeed, the marginal distribution given by $(\mathbf{c}_{1,1}, \ldots, \mathbf{c}_{1,M})$ and $(w_{1,1}, \ldots, w_{1,M})$ represents the first response-specific latent structure among groups, while the marginal distribution given by $(\mathbf{c}_{2,1}, \ldots, \mathbf{c}_{2,K})$ and $(w_{2,1}, \ldots, w_{2,K})$ represents the second response-specific one. The estimation of the parameters $\boldsymbol{B}$, $(\mathbf{C}_{11}, \ldots, \mathbf{C}_{MK})$, $(w_{11}, \ldots, w_{MK})$ and $\boldsymbol{\Sigma}$ is performed through the maximization of the likelihood function, mixture by the discrete distribution of random coefficients,

$$
\begin{aligned}
L(\mathbf{w}, \boldsymbol{B}, \mathbf{C}, \boldsymbol{\Sigma}|\mathbf{y}) = \sum_{m=1}^{M} \sum_{k=1}^{K} & \frac{w_{mk}}{\sqrt{|det(2\pi\boldsymbol{\Sigma})|^{J}}} \times \\
& \times \exp\left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} -\frac{1}{2} \begin{pmatrix} y_{1,ij} - c_{1,1m} - \sum_{p=1}^{P} \beta_{1p} x_{1p,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \sum_{p=1}^{P} \beta_{2p} x_{2p,ij} - c_{2,2k} z_{2,ij} \end{pmatrix}^{T} \boldsymbol{\Sigma}^{-1} \right. \\
& \left. \times \begin{pmatrix} y_{1,ij} - c_{1,1m} - \sum_{p=1}^{P} \beta_{1p} x_{1p,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \sum_{p=1}^{P} \beta_{2p} x_{2p,ij} - c_{2,2k} z_{2,ij} \end{pmatrix} \right\} \qquad (4)
\end{aligned}
$$

with respect to $\boldsymbol{B}$, the distribution of the random coefficients $(\mathbf{C}_{mk}, w_{mk})$, for $m = 1, \ldots, M$ and $k = 1, \ldots, K$, and $\boldsymbol{\Sigma}$, respectively.

One of the main novelty of this algorithm with respect to similar existent algorithms (see, for example, Aitkin (1996, 1999); Muthén (2004)) is that it does not fix *a priori* the number of support points of the random effects distribution, but it estimates it depending on the variability within the data. Namely, during the iterations of the EM algorithm, we implement a support points collapsing system in which two mass points closer than a fixed tolerance value D (in terms of euclidean distance) collapse to a unique point. This approach allows to deal with the identification of subpopulations from a different point of view with respect to methods that select the number of subpopulations based on the Dirichelet process or on the maximization of the likelihood *a posteriori* (Aitkin 1999; Dahl 2006). The threshold distance D is a tuning parameter that is related to the heterogeneity across subpopulations. Its value can be chosen standing on a rationale driven by the application and by the data values range. Appendix A reports details on how to assess the uncertainty of classification of the method, given the value D.

It is worth noticing that the bivariate modelling allows to estimate the association between the random effects relative to the two response variables. With Gaussian random effects, the association is measured by the Pearson's correlation coefficient. Here, with discrete random effects, the association can be estimated by looking at the frequencies in the matrix of the joint weights $\mathbf{w}$. In particular, we test the dependence of the two marginal distributions by means of the Pearson's chi-squared test and we estimate a measure of the association by computing the Cramer's V relative to the test

(Cramér 1999). Moreover, taking into account the support points values of the joint distribution, it is possible to compute the correlation between the two 2-dimensional (intercept and slope) vectors of random effects to investigate the correlation between the values of the support points relative to the two response variables (Puccetti 2019).

3 Simulation study

In this section, we test the performance of the BSPEM algorithm simulating nine situations in which the two response variables are related to each other in nine different ways, facing both structural correlation/uncorrelation between the subpopulations distributions and correlation/uncorrelation between the errors of the linear model.

We generate 10,000 bivariate observations that are nested within 100 groups in the following way:

$$\begin{aligned}
&\begin{pmatrix}\mathbf{y}_{1,i}\\ \mathbf{y}_{2,i}\end{pmatrix}^T = \mathbf{1}_{n_i}\begin{pmatrix}c_{1,1m}\\ c_{2,1k}\end{pmatrix}^T + \mathbf{x}_i\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}^T + \mathbf{z}_i\begin{pmatrix}c_{1,2m}\\ c_{2,2k}\end{pmatrix}^T + \boldsymbol{\epsilon}_i\\
&i = 1,\ldots,100 \qquad m = 1,\ldots,M \qquad k = 1,\ldots,K \\
&\boldsymbol{\epsilon}_i^T = \begin{pmatrix}\boldsymbol{\epsilon}_{1,i}\\ \boldsymbol{\epsilon}_{2,i}\end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind.
\end{aligned} \tag{5}$$

in which we set $M = 3$ and $K = 2$. We set $n_i = 100$, for $i = 1,\ldots,100$, and we make the following choice of parameters[5] $\mathbf{C}_{mk}$, for $m = \{1, 2, 3\}$ and $k = \{1, 2\}$:

Besides the coefficients, we sample the observations of the variables x, z and $\boldsymbol{\epsilon}$ as standard normal variables[6]:

$$\begin{aligned}
z_i &\sim \mathcal{N}(0, 1) \quad i = 1,\ldots,33\\
z_i &\sim \mathcal{N}(0, 1) \quad i = 34,\ldots,66\\
z_i &\sim \mathcal{N}(0, 1) \quad i = 67,\ldots,100
\end{aligned} \tag{6}$$

$$\begin{aligned}
x_i &\sim \mathcal{N}(0, 1) \quad i = 1,\ldots,33\\
x_i &\sim \mathcal{N}(0, 1) \quad i = 34,\ldots,66\\
x_i &\sim \mathcal{N}(0, 1) \quad i = 67,\ldots,100
\end{aligned} \tag{7}$$

and

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}_2\left(\mathbf{0}, \boldsymbol{\Sigma} = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right) \quad i = 1,\ldots,100. \tag{8}$$

Since we choose three different sets of parameters $(\mathbf{C}, \boldsymbol{B})$ to generate the data of the first response and two different sets to generate the ones of the second response, the data

[5] Note that this choice of parameters is finalized to the simulation study and it is driven only from the aim of a simple and clear visualization of the results. Any other choice of parameters is possible. Moreover, we consider the case of only one fixed covariate, but the all the considerations hold for any number of fixed covariates $P > 1$.

[6] In order for the metric to be consistent and for identifiabiliy issues, it is important to include only standardized covariates. Variables $\mathbf{x}$ and $\mathbf{z}$ are allowed to be different between first and second response variables (i.e. $\mathbf{x}_{1,i} \neq \mathbf{x}_{2,i}$).

Table 1 Set of parameters used in Eq. (5) to simulate data

	First response parameters	Second response parameters
$i = 1, \ldots, 33$	$c_{1,11} = 5$ $c_{1,21} = 10$ $\beta_1 = 3$	$c_{2,11} = 3$ $c_{2,21} = 1$ $\beta_2 = 2$
$i = 34, \ldots, 66$	$c_{1,12} = 2$ $c_{1,22} = 5$ $\beta_1 = 3$	$c_{2,11} = 3$ $c_{2,21} = 1$ $\beta_2 = 2$
$i = 67, \ldots, 100$	$c_{1,13} = 0$ $c_{1,23} = -2$ $\beta_1 = 3$	$c_{2,12} = 0$ $c_{2,22} = -3$ $\beta_2 = 2$

The intercepts and the coefficients of **z** differ across subpopulations, while the coefficients β of x are fixed. Colours highlight the different subpopulations related to each response variable. We impose a structure with three subpopulations in the first response (M = 3) and two subpopulations in the second one (K = 2)

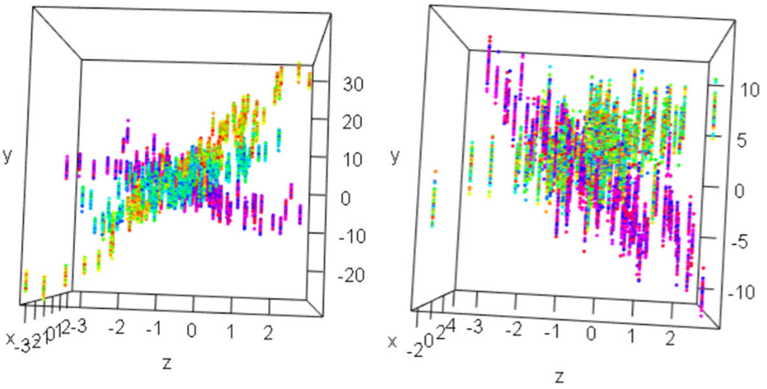


Fig. 1 Data simulated with the set of parameters reported in Table 1 and values of **x**, **z** and $\boldsymbol{\epsilon}$ defined in Eqs. (6), (7) and (8) respectively. Figure on the left panel represents the first response and figure on the right panel represents the second one. It is possible to identify the presence of three and two subpopulations in the first and in the second response respectively. Colors are automatically assigned by the software R

related to the first response are clustered within three subpopulations (M=3), while the ones related to the second one are clustered within two subpopulations (K=2). Figure 1 shows the data simulated with the set of parameters reported in Table 1.

The correlation among the two response variables depends both on the subpopulations distributions that we use to generate them (i.e. on the choice of $\mathbf{C}_{mk}$) and on the correlation between the errors. In this perspective, the parameters distribution shown in Table 1 induces a structural correlation among the subpopulations of the two response variables, since the bivariate distribution of $\mathbf{C}_{mk}$ follows a precise structure among the groups. Regarding the distribution of the errors, the covariance of the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ in Eq. (8) is set to zero, implying the absence of any further correlation among the two responses.

We apply the BSPEM algorithm to this simulated dataset, choosing $D = 1$ and $\texttt{tollR} = \texttt{tollF} = 10^{-2}$ (see Algorithm 1 in Appendix A). We repeat the simulation for 100 runs. On average, the algorithm converges in 6 iterations and it always identifies

Table 2 Values of the parameters of Eq. (5) estimated by the BSPEM algorithm, obtained as the average over the 100 runs (for each parameter we also report its Mean Square Error in brackets)

	First response parameters	Second response parameters
$i = 1, \ldots, 33$	$\hat{c}_{1,11} = 5.00085$	$\hat{c}_{2,11} = 3.01097$
	$(MSE_{1,11} = 0.00024)$	$(MSE_{2,11} = 0.00024)$
	$\hat{c}_{1,21} = 9.99876$	$\hat{c}_{2,21} = 1.00384$
	$(MSE_{1,21} = 0.00028)$	$(MSE_{2,21} = 0.00091)$
	$\hat{\beta}_1 = 2.99856$	$\hat{\beta}_2 = 1.99854$
	$(MSE_{\beta_1} = 0.00059)$	$(MSE_{\beta_2} = 0.00065)$
$i = 34, \ldots, 66$	$\hat{c}_{1,12} = 2.01128$	$\hat{c}_{2,11} = 3.01066$
	$(MSE_{1,12} = 0.00037)$	$(MSE_{2,11} = 0.00024)$
	$\hat{c}_{1,22} = 4.99942$	$\hat{c}_{2,21} = 1.01334$
	$(MSE_{1,22} = 0.00024)$	$(MSE_{2,21} = 0.00091)$
	$\hat{\beta}_1 = 2.99856$	$\hat{\beta}_2 = 1.99854$
	$(MSE_{\beta_1} = 0.00059)$	$(MSE_{\beta_2} = 0.00065)$
$i = 67, \ldots, 100$	$\hat{c}_{1,13} = 0.00068$	$\hat{c}_{2,12} = -0.00768$
	$(MSE_{1,13} = 0.00195)$	$(MSE_{2,12} = 0.00065)$
	$\hat{c}_{1,23} = -2.00531$	$\hat{c}_{2,22} = -2.99967$
	$(MSE_{1,23} = 0.00203)$	$(MSE_{2,22} = 0.00182)$
	$\hat{\beta}_1 = 2.99856$	$\hat{\beta}_2 = 1.99854$
	$(MSE_{\beta_1} = 0.00059)$	$(MSE_{\beta_2} = 0.00065)$

Colors represent the different subpopulations identified by the algorithm. The algorithm identifies three subpopulations (M = 3) for the first response and two subpopulations for the second one (K = 2)

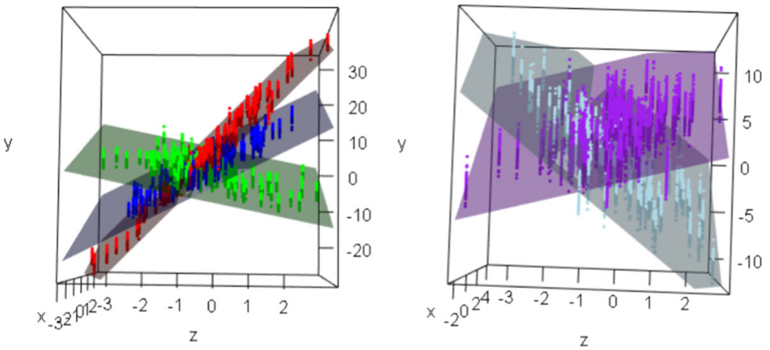


Fig. 2 Simulated data with the regression planes identified by the BSPEM algorithm in one of the 100 runs. Colors represent the different subpopulations: three for the first response (figure on the left panel) and two for the second response (figure on the right panel). The estimated parameters of the regression planes are shown in Table 2

the correct number of clusters for both the two response variables, whose estimated parameters (mean and MSE over the 100 runs) are shown in Table 2.

Figure 2 shows the data with the regression planes identified by the algorithm in one of the 100 runs, for both the two response variables.

The algorithm assigns each group i, for $i = 1, \ldots, 100$, to the correct cluster related to the two response variables, that means that assigns each group i, for $i = 1, \ldots, 100$, to the correct bivariate cluster (m, k). The estimates of the $(M \times K)$-dimensional matrix of weights $\mathbf{w}$ and of $\boldsymbol{\Sigma}$, averaged over the 100 runs, are the following:

$$\hat{\mathbf{w}} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix} \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.0012 & 0.0022 \\ 0.0022 & 0.9996 \end{pmatrix} \tag{9}$$

$$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0002 & 0.0001 \\ 0.0001 & 0.0002 \end{pmatrix}. \tag{10}$$

By looking at the matrix $\hat{W}$, we can identify the distribution of the groups on the support, composed by the 6 mass points. Since we impose a structural correlation between the clusters distribution of the two response variables (see the coefficients in Table 1), the estimated distribution of the weights w_{mk} is not uniform on the $M \times K$ masses, but it is possible to recognize the pattern that we used to generate the data. Regarding the variance/covariance matrix $\hat{\boldsymbol{\Sigma}}$, the covariance is correctly estimated as null and the two estimated variances are also close to 1.

The case just shown represents the particular situation in which the subpopulations distributions are not uniform on the mass points and the errors are not correlated, but it can also be the case that the two response variables do not present correlated subpopulations or even present correlated errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$. In order to test the performance of the BSPEM algorithm in these further cases, we modify the values of $\mathbf{C}_{mk}$ and $\boldsymbol{\epsilon}$ in order to simulate nine different scenarios. The simulated scenarios result from the intersection of three different assumptions both on the structural correlation among subpopulations and on the dependence structure of the errors. In particular,

- Latent subpopulations structure:
  1. Structural correlation among subpopulations relative to the two response variables (i.e. maximum dependence in the weights matrix);
  2. Partial structural correlation among subpopulations relative to the two response variables;
  3. No structural correlation among subpopulations relative to the two response variables (i.e. independence in the weights matrix);
- Dependence between the errors:
  1. Dependence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ with correlation coefficient $\rho = 1$;
  2. Dependence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ with correlation coefficient $\rho = 0.5$;
  3. Independence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ with correlation coefficient $\rho = 0$.

In order to avoid any type of structural correlation among the subpopulations of the two response variables, i.e. in order to have a subpopulations distribution uniform on the mass points, we randomly shuffle the order of the parameters shown in Table 1 across the 100 groups, so that there are no definite patterns on the parameters $\mathbf{C}_{mk}$ between the two responses. For the case of partial structural correlation, we shuffle only part of the groups. In particular, we maintain the first 33 groups as shown in Table 1, while we shuffle the remaining 67 ones. In order to simulate the dependence/independence among the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$, we set the variance/covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 0.51 & 0.5 \\ 0.5 & 0.51 \end{pmatrix}$ for the first case ($\rho = 1$), $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ for the second case ($\rho = 0.5$) and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for the third case ($\rho = 0$).

We apply the BSPEM algorithm to these four different types of simulated data (100 runs for each of the four cases), with the same choice of parameters $D = 1$, `tollR` = `tollF` = 10^{-2} (see Algorithm 1 in Appendix A). The algorithm is able to identify the correct subpopulations distribution in all the nine situations. The visualization of the results in all the nine cases is similar to the one shown in Fig. 2 and the estimates of the parameters $\mathbf{C}_{mk}$, for $m = 1, \ldots, 3$ and $k = 1, 2$ and $\boldsymbol{B}$ in the nine cases are in line with the ones shown in Table 2. The differences across the nine cases are the estimates of the weights matrices $\mathbf{w}$ and of $\boldsymbol{\Sigma}$, whose means over the 100 runs are shown in Table 3.

From Table 3, we see that the model is identifiable, since it is able to distinguish the correlation among the two response variables that is given by a structural correlation among subpopulations distribution (showed in $\mathbf{w}$) from the correlation imposed by dependent errors (showed in $\boldsymbol{\Sigma}$). In the last column of Table 3, where we do not impose any structural correlation among subpopulations, the distribution of the weights, less than small variations, is uniform on the mass points. In the second column of the table, where we impose a partial structural correlation among the two subpopulations distributions, we observe that the 33% of groups belongs to subpopulation (1,1), while the remaining 67% is uniformly distributed on the other support points.

Finally, since in the case study data have a group size ranging from 10 to 28, with a mean of about 17, we add a further check repeating the first simulation, but considering $n_i = 20$ instead of $n_i = 100$, for $i = 1, \ldots, 100$. Appendix B reports the results that confirm that the method is robust with respect to group sizes.

The only parameter that significantly influences the results of the BSPEM algorithm is the threshold distance D (see Algorithm 1 in Appendix A). In order to give an idea of the sensitivity of the algorithm to the values of D, in the cases seen above, the algorithm gives the same result for each value of D between 0.5 and 2. For values of $D < 0.5$, the BSPEM algorithm is too sensitive to the variability among the data and identifies more that 6 mass points, while for values of $D > 2$, the algorithm does not entirely catch the variability among the data identifies less than 6 mass points.[7]

4 Case study: application to Italian middle schools (grades 6–8)

In this section, we present our dataset, that deals with a sample of Italian middle schools in 2016/2017. We apply the BSPEM algorithm to identify subpopulations of classes, on the basis of their different effects on mathematics and reading student achievements.

The sample that we consider is composed by students and classes that take the INVALSI test under the supervision of the INVALSI staff. This sample regards the 10% of the total population and it is directly selected by INVALSI in order to be

[7] Further information regarding the choice of the threshold value D is given in Masci et al. (2019).

Table 3 Estimates of the weights matrix $\mathbf{w}$ and of the variance/covariance matrix $\boldsymbol{\Sigma}$ (with its MSE computed over the 100 runs) of model in Eq. (5) for the nine different cases of values of $\mathbf{C}_{mk}$ and $\boldsymbol{\epsilon}$

	Structural correlation among subpopulations	Partial structral correlation among subpopulations	No structural correlation among subpopulations
$\epsilon_1 \not\perp \epsilon_2$	$\hat{\mathbf{w}} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix}$	$\hat{\mathbf{w}} = \begin{pmatrix} 0.33 & 0.00 \\ 0.14 & 0.19 \\ 0.19 & 0.15 \end{pmatrix}$	$\hat{\mathbf{w}} = \begin{pmatrix} 0.25 & 0.08 \\ 0.21 & 0.12 \\ 0.20 & 0.14 \end{pmatrix}$
$\boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.5006 & 0.5006 \\ 0.5006 & 0.5007 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.5056 & 0.5054 \\ 0.5054 & 0.5055 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.4999 & 0.4998 \\ 0.4998 & 0.4999 \end{pmatrix}$
	$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0003 & 0.0001 \\ 0.0001 & 0.0004 \end{pmatrix}$	$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0002 & 0.0001 \\ 0.0001 & 0.0004 \end{pmatrix}$	$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0003 & 0.0002 \\ 0.0002 & 0.0001 \end{pmatrix}$
$\epsilon_1 \not\perp \epsilon_2$	$\hat{\mathbf{w}} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix}$	$\hat{\mathbf{w}} = \begin{pmatrix} 0.33 & 0.00 \\ 0.15 & 0.18 \\ 0.08 & 0.16 \end{pmatrix}$	$\hat{\mathbf{w}} = \begin{pmatrix} 0.23 & 0.10 \\ 0.22 & 0.12 \\ 0.21 & 0.12 \end{pmatrix}$
$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.0162 & 0.5011 \\ 0.5011 & 0.9697 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.0076 & 0.4952 \\ 0.4952 & 0.9979 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.9933 & 0.4887 \\ 0.4887 & 1.0234 \end{pmatrix}$
	$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0002 & 0.0001 \\ 0.0001 & 0.0004 \end{pmatrix}$	$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0003 & 0.0001 \\ 0.0001 & 0.0003 \end{pmatrix}$	$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0004 & 0.0002 \\ 0.0002 & 0.0003 \end{pmatrix}$
$\epsilon_1 \perp\!\!\!\perp \epsilon_2$	$\hat{\mathbf{w}} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix}$	$\hat{\mathbf{w}} = \begin{pmatrix} 0.33 & 0.00 \\ 0.17 & 0.16 \\ 0.16 & 0.18 \end{pmatrix}$	$\hat{\mathbf{w}} = \begin{pmatrix} 0.23 & 0.10 \\ 0.22 & 0.12 \\ 0.21 & 0.12 \end{pmatrix}$
$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.0163 & 0.0015 \\ 0.0015 & 0.9691 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.9886 & -0.0016 \\ -0.0016 & 0.9978 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.9936 & 0.0094 \\ 0.0094 & 1.0223 \end{pmatrix}$
	$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0002 & 0.0001 \\ 0.0001 & 0.0004 \end{pmatrix}$	$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0005 & 0.0002 \\ 0.0002 & 0.0004 \end{pmatrix}$	$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0005 & 0.0001 \\ 0.0001 & 0.0003 \end{pmatrix}$

Table 4 Student level variables of the INVALSI database 2016/2017 used in the analysis with their descriptive statistics

Variable	Type	Mean	sd	Median	IQR
`Math8`	Cont	53.201	20.036	52.489	29.322
`Read8`	Cont	64.491	17.278	66.392	23.001
`Math5`	Cont	68.475	16.641	70.000	26.001
`Read5`	Cont	66.608	16.736	68.965	24.138
`ESCS`	Cont	0.147	0.991	0.069	1.323
`Gender`	0/1	0.51	–	–	–
`Immig`	0/1	0.07	–	–	–

representative of the entire Italian population. Being the test in this sample supervised by the INVALSI staff, we overcome the potential problems related to the cheating of students or teachers. We restrict the sample to classes with at least 10 students. The sample comprises 18,242 students nested within 1082 classes.[8]

4.1 The database about the Italian middle schools

The database includes data about students attending grade III of junior secondary school in year 2016/2017. About these students, besides their results of the INVALSI tests in reading and mathematics at grade 8 (`read8` and `math8` respectively), we consider other five variables: the INVALSI test scores in reading and mathematics of these students three years before, i.e. at the last year of primary school (`read5` and `math5` respectively); the socioeconomic index (`ESCS`) that is an index built by INVALSI by considering parents' occupation and educational titles and the possession of certain goods at home (for instance, computer or the number of books); the gender of the student (`gender`, $1 = $ female, $0 = $ male) and the immigrant status (`immig`, $0 = $ Italian, $1 = $ first/second generation immigrant).[9] The INVALSI test score is a continuous variable that takes values between 0 and 100 (proportion of correct answers in the test), while the ESCS is built as a continuous variable with mean equal to 0 and variance equal to 1. Controlling for prior achivement at grade 5 allows the model to be specified as a value-added. Table 4 reports the five student level variables used in the analysis with their descriptive statistics[10]. In the considered cohort of students, 51% are females and 7% are not native Italians, but 1^{st} or 2^{nd} generation immigrants. On average, the INVALSI test scores are slightly higher at grade 5 than at grade 8 (we deal with this factor by standardizing values, see Sect. 4.2).

In 2016/2017, INVALSI collected information about classes and teachers by means of a dedicated questionnaire. This questionnaire includes an abundant set of information about the class body composition, the approach of the teacher to INVALSI tests,

[8] We remind that in our sample each class is within a different school, i.e. we do not observe more classes in the same school.

[9] The choice of these variables relies on the fact that the literature identifies them as significant for predicting student achievements, cross-sectionally to many studies (see, among the others, Masci et al. (2017); Agasisti et al. (2017); Agasisti and Vittadini (2012)).

[10] In the analysis, these variables will be standardized.

Table 5 Teacher and class levels variables of the INVALSI database 2016/2017 used in the analysis with their explanation

Variable	Type	Explanation
Teachers general questions (for both maths and reading teachers)		
`Updated techniques`	y/n	The teacher applies new techniques learned at refreshment courses
`Team work or research`	y/n	The teacher organizes team work or research in groups for students
`Extra activities`	y/n	The teacher organizes extra scholastic activities for student reinforcement
`Computer/internet`	y/n	The teacher uses media support in class
`Refresher courses`	Num	Number of refreshment courses the teacher had in the last two years
`Contacts among teachers`	y/n	Teacher exchanges views with other teachers
Teachers personal information (for both maths and reading teachers)		
`Num years of teaching here`	$1:4$	Since how many years the teacher teaches in the actual school. 1: one year or less; 2: 2-3 years; 3: 4-5 years; 4: > than 5 years.
`Permanent job`	y/n	The teacher has a permanent contract
`Gender`	y/n	y= male; n = female.
`Age`	Num	Age of the teacher
`Education`	$1:3$	Higher level of education of the teacher 1: less than degree; 2: degree; 3: phd/master

personal information of the teacher (age, education, gender), teaching practices and available materials in the class. Tables 5, 6, 7 report teacher and class level variables that we consider, following suggestions derived from the literature about school effectiveness (David et al. 2000), with their explanation.

The variables shown in Table 7 cover the four areas that regard (i) the class body composition, (ii) teacher personal information (gender, age, education, …), (iii) teaching practices of the teacher and (iv) teacher's perception about the work and the collaboration within the school and about the school principal. Class body composition and teacher personal information have been broadly considered in the literature as potential influencer of student learning (Palardy 2008; Winkler 1975; Dar and Resh 1986, 2018; Belfi et al. 2012; Wayne and Youngs 2003). More recent studies investigate also the effets of different teaching approaches (traditional versus modern teaching methods) on student learning, finding heterogeneous results (Brewer and Goldhaber 1997; Schwerdt and Wuppermann 2011; Bietenbeck 2014; De Witte and Van Klaveren 2014; Wenglinsky 2002). Therefore, besides information regarding the class body composition, the geographical area and personal information of the teacher, we decided to select from the questionnaire the information that describes the type of teaching method of the teacher (i.e. the student skills that the teacher stress more and aim to develop, the type of exercises that the teacher does in class and the type of tests that the teacher prepares for students) and the managerial practices adopted by the school principal.

Table 6 Teacher and class levels variables of the INVALSI database 2016/2017 used in the analysis with their explanation

Questions about school principals (for both maths and reading teachers)		
Princ refreshment courses	y/n	The school principal encourages teachers to follow refreshment courses
Princ lineup teach	y/n	The school principal organizes lineup meetings for teachers
Princ evaluate	y/n	The school principal evaluates the teachers in their job
Only for mathematics teachers		
num mathematics hours	Num	Number of hours of maths lesson per week
Main teaching method	cat	‘a’: teach definitions and theorems that students can apply to solve new problems
		‘b’: Favor the maths language and the capacity of using formulas written in symbols
		‘c’: Favor meanings of maths symbols
		‘d’: Favor the capacity of build concepts, models and theories
Oral individ exam	y/n	The teacher tests students by means of oral individual exams
Oral group exam	y/n	The teacher tests students by means of oral exams for groups of students
Teacher written exam	y/n	The teacher tests students by means of written exam made by him/herself
Book written exam	y/n	The teacher tests students by means of written exam taken by the book
Calculations alone	y/n	The teacher teaches students to make calculations without the support of the calculator
Table diagram graph	y/n	The teacher teaches students to interpret tables, diagrams and graphs
Maths memory	y/n	The teacher asks students to memorize maths rules and theorems
Graphs for problems	y/n	The teacher teaches students to analyze graphs to solve maths problems

Table 7 Teacher and class levels variables of the INVALSI database 2016/2017 used in the analysis with their explanation

Variable	Type	Explanation
Only for reading teachers		
`Num reading hours`	Num	Number of hours of reading lesson per week
`Programmed oral exam`	y/n	The teacher tests students by means of programmed oral exam
`Not programmed oral exam`	y/n	The teacher tests students by means of not programmed oral exam
`Grouped oral exam`	y/n	The teacher tests students by means of oral exam for groups of students
`Teacher close test`	y/n	The teacher tests students by means of written close questions tests made by him/herself
`Teacher open test`	y/n	The teacher tests students by means of written open questions tests made by him/herself
`Teacher book test`	y/n	The teacher tests students by means of written tests taken by the book
`Summarize text`	y/n	The teacher trains students to summarize texts
`Write reflections`	y/n	The teacher trains students to write texts about their reflections and thinking
`Read newspaper`	y/n	The teacher trains students to read newspapers and journals
Class information and body composition		
`Area geo`	Cat	Northern/Central/Southern Italy
`Nstud`	Num	Number of students
% `Stud antic`	Num	Percentage of early-enrolled students
% `Stud postic`	Num	Percentage of late-enrolled students
% 1^{st}`-gen immig`	Num	Percentage of first generation immigrants
% 2^{nd}`-gen immig`	Num	Percentage of second generation immigrants

4.2 BSPEM applied to data of Italian middle schools: estimating subpopulations of classes

The semi-parametric two-level linear model applied to INVALSI data, considering students (level 1) nested within classes (level 2), takes the following form:

$$\begin{aligned}
\mathbf{Y}_i &= \mathbf{1}_{n_i}\begin{pmatrix} c_{1,1m} \\ c_{2,1k}\end{pmatrix}^T + \sum_{p=1}^{P}\mathbf{x}_{ip}\begin{pmatrix}\boldsymbol{\beta}_{1p} \\ \boldsymbol{\beta}_{2p}\end{pmatrix}^T + \mathbf{z}_i\begin{pmatrix} c_{1,2m} \\ c_{2,2k}\end{pmatrix}^T + \boldsymbol{\epsilon}_i \\
& i = 1,\ldots,N \quad m = 1,\ldots,M, \quad k = 1,\ldots,K \qquad (11)\\
\boldsymbol{\epsilon}_i^T &= \begin{pmatrix}\boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i}\end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind.
\end{aligned}$$

where i is the class index and N is the total number of classes. $\mathbf{Y}_i = \big(\texttt{math8}_i \quad \texttt{read8}_i\big)$ is the bivariate vector of the INVALSI test scores of students attending grade 8, in mathematics and reading. $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ is the $(n_i \times 3)$-matrix of the fixed covariates, that comprehends socioeconomic index, gender and immigrant status. $\mathbf{z}$ is the vector of INVALSI test score of the same students but three years before (at grade 5), that differs across the two response variables, being `math5` for the first response (`math8`) and `read5` for the second one (`read8`). In particular, we standardize the variables `math8`, `read8`, `math5`, `read5` and `ESCS`, so that they all have mean equal to 0 and variance equal to 1. Our interest is to see how the association between the INVALSI test score at the end of the primary school/beginning of the junior secondary school and the INVALSI test score at the end of the junior secondary school does change across students attending different classes, after adjusting for some student level confounding factors (socioeconomic index, gender and immigrant status), both in reading and mathematics. The period between grade 5 and grade 8 is the entire period of the junior secondary school and this association represents a kind of class effect, seen as the impact that the class has on the evolution of its student achievements. With this modeling, we identify subpopulations of classes within which class impacts are similar and across which they are different. The bivariate nature of the modeling allows to do that both for reading and mathematics achievements, considering also the joint effect of the class on the two school subjects. We apply the BSPEM algorithm with the following choice of parameters: $D_1 = D_2 = 0.3$, $\tilde{w}_1 = \tilde{w}_2 = 0.01$, $\texttt{tollR} = \texttt{tollF} = 10^{-2}$, $\texttt{it} = 40$, `itmax=20`, `it1=20` (see Algorithm 1 in Appendix A).[11] The algorithm converges in 30 iterations and identifies $M = 5$ mass points for the random coefficients distribution related to the first response (mathematics) and $K = 4$ mass points for the one related to the second response (reading). From an educational viewpoint, for interpretation, classes can be classified into five homogeneous groups

[11] We choose $\tilde{w}_1 = \tilde{w}_2 = 0.01$ in order to observe subpopulations of classes containing at least 100 classes. Our interest is in the identification of relevant trends that describe most of the population and that can be characterized *a posteriori* in terms of class- and school-level variables. A lower value of $\tilde{w}$ is possible and it allows to identify smaller subpopulations composed by what we interpret as outlier classes. The choice of the threshold distances D_1 and D_2 is driven by the entropy of the conditional weights matrices W_1 and W_2. We choose for D_1 and D_2 the lowest values that allow to maintain low entropy values (see Appendix A for details about the computation of the entropy).

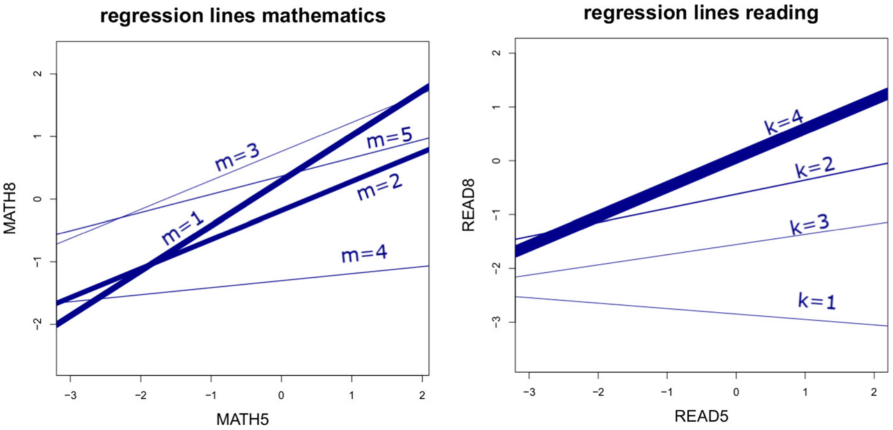


Fig. 3 Regression planes projected on the 2-dimensional plane identified by the answer variable and the random covariate, identified by the parameters of Eq. (11) estimated by the BSPEM algorithm and whose parameters are shown in Table 8. Panel on the left reports the results for the first response, while panel on the right reports the results for the second one. The algorithm identifies $M = 5$ mass points for the first response and $K = 4$ mass points for the second one. For a better visualization, we do not represent all the observations but only the identified regression lines. Line widths are proportional to the marginal weights w_1 and w_2

when considering value-added in mathematics, while in four groups when considering value-added in reading. The estimates of the identified parameters (which measure the effectiveness of classes) are shown in Table 8.

$\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are the fixed coefficients and therefore their estimates are stable across the subpopulations; $\hat{\mathbf{c}}_{1,m}$, for $m = 1, \ldots, 5$ and $\hat{\mathbf{c}}_{2,k}$, for $k = 1, \ldots, 4$ are the estimates of the random coefficients and $\hat{w}_1$ and $\hat{w}_2$ are the estimated weights related to the marginal distributions of the two random effects. Regarding the fixed effects (i.e. the individual-level covariates that affect students' performance), the positive coefficient of the variable ESCS (0.089 for mathematics and 0.095 for reading) suggests that students with higher ESCS are associated to a higher progress between grade 5 and grade 8 INVALSI scores; females have on average higher scores in reading and lower ones in mathematics, with respect to males (coefficient of gender is −0.055 for mathematics and 0.219 for reading); being an immigrant student has a negative effect in reading, but a slightly positive one in mathematics, once controlling for other individual characteristics (coefficient of immigrant is −0.083 for reading and 0.048 for mathematics). In order to visualize the results related to random effects (class effectiveness), Fig. 3 reports the regression planes identified for both the two response variables, projected on the 2-dimensional plane identified by the answer variable and the random covariate.

By looking at the estimated parameters in Table 8 and the regression lines in Fig. 3, it is possible to make considerations about the identified subpopulations of classes. Such classification is particularly useful for decision-makers, who can have a clear image of the heterogeneous effect of attending classes with different characteristics. Among the five identified subpopulations related to the class effect in mathematics, subpopulation $m = 4$ (containing 6.4% of the classes) clearly contains the classes

Table 8 Estimates of the parameters of Eq. (11) obtained by the BSPEM algorithm, related to the two response variables

First response variable

	$\hat{c}_{1,1}$ (intercept)	$\hat{c}_{1,2}$ ($\texttt{math5}$)	$\hat{w}_1$ (weight)	$\hat{\beta}_{11}$ ($\texttt{ESCS}$)	$\hat{\beta}_{12}$ ($\texttt{gender}$)	$\hat{\beta}_{13}$ ($\texttt{immigrant}$)
m=1	0.295	0.719	0.458	0.089	−0.055	0.048
m=2	−0.181	0.464	0.384			
m=3	0.762	0.463	0.025			
m=4	−1.301	0.112	0.064			
m=5	0.366	0.291	0.069			

Second response variable

	$\hat{c}_{2,1}$ (intercept)	$\hat{c}_{2,2}$ ($\texttt{read5}$)	$\hat{w}_2$ (weight)	$\hat{\beta}_{21}$ ($\texttt{ESCS}$)	$\hat{\beta}_{22}$ ($\texttt{gender}$)	$\hat{\beta}_{23}$ ($\texttt{immigrant}$)
k=1	−2.848	−0.101	0.019	0.095	0.219	−0.083
k=2	−0.622	0.262	0.095			
k=3	−1.556	0.188	0.018			
k=4	0.054	0.544	0.868			

The coefficients $\boldsymbol{\beta}$ do not change across subpopulations with the worse effect on student achievements, since the predicted values of y are the lowest for almost the entire range of previous score `math5`. Subpopulations $m = 1$ and $m = 2$ (containing 45.8% and 38.4% of the classes, respectively) represent the most common trends and with respect to them, subpopulations $m = 3$ and $m = 5$ have the two following characteristics: subpopulation $m = 3$ (2.5% of the classes) can be interpreted as the best set of classes since the predicted values of y are the highest in almost the entire range of the covariate `math5`; subpopulation $m = 5$ (6.9% of the classes) contains classes where students have on average higher predicted values of INVALSI score at grade 8 than the ones in subpopulation $m = 2$, while with respect to population $m = 1$ they have higher predicted values of y for values of `math5` smaller than 0, while they have lower predicted values of y for values of `math5` bigger than 0. These subpopulations contain classes which exert heterogeneous effects on achievements, namely their effectiveness is different along the distribution of initial students' ability (as measured by test score at grade 5). Regarding the results of reading, the four identified subpopulations are very well distinct. The subpopulation of the worst classes corresponds to subpopulation $k = 1$ (containing about 2% of the classes), that is characterized by a very low intercept and a slightly negative slope: students attending classes that belong to this subpopulation have a low predicted value of INVALSI score, regardless of the fact that they had high or low scores at grade 5. On the opposite, subpopulation $k = 4$ (containing 86.8% of the classes) contains the set of the best classes since for all values of previous score z between -3 and 2, i.e. for almost the entire range of values of the random covariate, the predicted value of y is higher that the ones of the other subpopulations of classes. Subpopulation $k = 2$ (containing 9.5% of the classes) is the second one in terms of high values of predicted score y, while subpopulation $k = 3$ (containing 1.8% of the classes) have predicted values of y lower than the ones of subpopulations $k = 4$ and $k = 2$ but higher than the ones of subpopulation $k = 1$.

The algorithm also identifies the reference subpopulations, that are the most numerous ones, and the subpopulations that depart from them, composed by classes that have an exceptional effect, whether positive or negative.

The interpretations of these subpopulations are also supported by the average values of the standardized variables across them,[12] reported in Table 9. Regarding mathematics, subpopulation $m = 4$ contains classes where the average score of `math5` is the highest ($\overline{\texttt{math5}}_1 = 0.224$), but where the average score of `math8` is the lowest ($\overline{\texttt{math8}}_1 = -1.351$), confirming the negative effects (value-added) of the classes that belong to this subpopulation on students' achievement. Subpopulation $m = 3$, interpreted as the subpopulation containing classes with the highest positive effect, is characterized by the lowest average score of `math5` ($\overline{\texttt{math5}}_2 = -0.118$), but with the highest average score of `math8` ($\overline{\texttt{math8}}_2 = 0.753$). This subpopulation is the one with the highest average student `ESCS`. When considering reading, subpopulation $k = 4$, interpreted as the one containing the best classes, is indeed characterized by the lowest average value of `read5` ($\overline{\texttt{read5}}_1 = -0.051$) and the highest average score of `read8` ($\overline{\texttt{read8}}_1 = 0.138$). Also in this case, this subpopulation is characterized by

[12] These average values are obtained by computing the means of the variables over all students attending classes that belong to the different subpopulations.

Table 9 Average values of some student level variables used in the analysis, across the identified subpopulations (five for mathematics and four for reading)

First response variable			
	$\overline{\texttt{math8}}$	$\overline{\texttt{math5}}$	$\overline{\texttt{ESCS}}$
m=1	0.259	−0.049	0.103
m=2	−0.214	0.007	−0.106
m=3	0.753	−0.118	0.102
m=4	−1.351	0.224	−0.432
m=5	0.326	−0.075	−0.078
Second response variable			
	$\overline{\texttt{read8}}$	$\overline{\texttt{read5}}$	$\overline{\texttt{ESCS}}$
k=1	−2.78	0.427	−0.075
k=2	−0.518	0.149	−0.345
k=3	−1.398	0.342	−0.128
k=4	0.138	−0.051	0.014

the highest average value of ESCS. On the other side, subpopulation $k = 1$, associated to a negative class effect, has the highest average value of read5 ($\overline{\texttt{read5}}_4 = 0.427$) and the lowest average value of read8 ($\overline{\texttt{read8}}_4 = -2.78$).

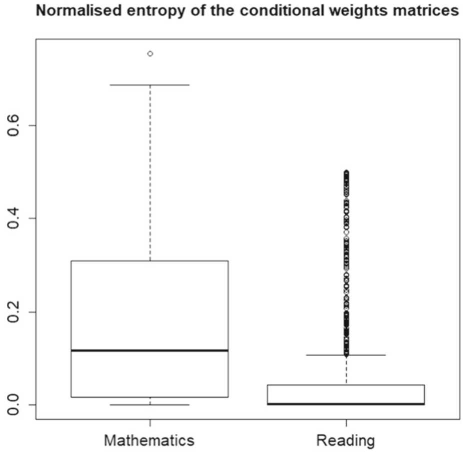
The $M \times K$ matrix of the joint weights $\mathbf{w}$ and the variance/covariance matrix $\boldsymbol{\Sigma}$ are estimated as follows:

$$\hat{\mathbf{w}} = \begin{pmatrix} 0.0000 & 0.0007 & 0.0003 & 0.4571 \\ 0.0054 & 0.0518 & 0.0047 & 0.3220 \\ 0.0022 & 0.0000 & 0.0023 & 0.0204 \\ 0.0068 & 0.0312 & 0.0082 & 0.0179 \\ 0.0043 & 0.0111 & 0.0029 & 0.0507 \end{pmatrix} \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.455 & 0.183 \\ 0.183 & 0.451 \end{pmatrix}.$$

The covariance and the correlation among the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are 0.183 and 0.404, respectively. We now focus on the distributions of groups on the bivariate subpopulations (weights matrix $\hat{\mathbf{w}}$) in order to estimate the association between the two random effects distributions. The Pearson chi-squared test of independence produces a p-value lower than $2.2e^{-16}$, rejecting the hypothesis of independence. The Cramér's V takes value 0.3176. This result suggests that there is a pattern in the distribution of groups within the subpopulations relative to reading and mathematics. The dependence between the two distributions supports the importance of the bivariate modelling. In order to investigate the correlation between the two 2-dimensional vectors of support point, we compute the correlation coefficients between random vectors (Puccetti 2019), as follows:

$$\rho(\mathbf{C}_1, \mathbf{C}_2) = \frac{\operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{C}_1\mathbf{C}_2})}{\operatorname{tr}((\boldsymbol{\Sigma}_{\mathbf{C}_1} \times \Sigma_{\mathbf{C}_2})^{1/2})}$$

Fig. 4 Normalised entropy distributions of the conditional weights matrices W_1 and W_2



where $\boldsymbol{\Sigma}_{\mathbf{C}_1}$ and $\boldsymbol{\Sigma}_{\mathbf{C}_2}$ are the variance/covariance matrices of random effects relative to the first and second responses, respectively; $\boldsymbol{\Sigma}_{\mathbf{C}_1\mathbf{C}_2}$ is the cross-covariance matrix whose elements $(\boldsymbol{\Sigma}_{\mathbf{C}_1\mathbf{C}_2})_{ij}$ are computed as $\text{Cov}(\mathbf{c}_{1i}, \mathbf{c}_{2j})$, for $i, j = 1, 2$. The correlation $\rho(\mathbf{C}_1, \mathbf{C}_2)$ results to be 0.3497. This low value of correlation reveals that the internal dynamics of a class in reading and mathematics might be substantially different. In the literature, authors have found different magnitude of class effects correlation between reading and mathematics [see, among the others, Grilli et al. (2016); Masci et al. (2017)]. In the comparison of this result with the ones obtained in the previous literature, we would like to remark that the inclusion of both random intercept and slope and, consequently, the use of Puccetti's correlation coefficient for 2-dimensional vectors instead of the classical Pearson's correlation coefficient might lead to unbalanced comparisons.

In order to measure the uncertainty of classification of the BSPEM method, we compute the entropy of the conditional weights matrices W_1 and W_2. In order to compare the results between mathematics and reading, we normalise the two entropy distributions with respect to their maxima (that are $-log(1/5) = 1.61$ for mathematics and $-log(1/4) = 1.38$ for reading). Figure 4 shows the distribution of the normalised entropy, computed on the rows of the two conditional weights matrices W_1 and W_2 (details in Appendix A).

The mean and the median of the two normalised entropy distributions are respectively 0.173 and 0.117 for mathematics and 0.059 and 0.002 for reading. On a scale between 0 and 1, where 0 corresponds to the minimum entropy and 1 to the maximum one, the low normalised entropy values of our case study confirm that the method assigns classes to the identified subpopulations with a low level of uncertainty, on average. In particular, the level of uncertainty is, on average, significantly lower for reading than for mathematics, suggesting that subpopulations of classes are better distinguished for reading than for mathematics.

Finally, we focus on the residual variation of the model, investigating it for our discrete random effects case. In multilevel models, the residual variation is split into component parts that are attributed to both student and class levels. The Variance

Partition Coefficient (VPC) measures the percentage of variation that is attributable to the highest-level (classes) sources of variation and conveniently summarizes the ‘importance’ of classes (Goldstein et al. 2002). The VPC is expressed as:

$$VPC = \frac{\tau}{\tau + \sigma^2} \tag{12}$$

where τ is the *between-classes* variance and σ^2 is the *within-classes* variance. In order to compute the *between-classes* variance, we need to explicate the variance/covariance matrix of random effects. Following the theory presented by Rights and Sterba (2016) about the relationship between parametric and non-parametric mixed-effects models, we compute the *implied* variance/covariance matrix Γ of random effects. In our random intercept + slope multilevel model, Γ is expressed as:

$$\Gamma = \begin{pmatrix} \mathrm{Var}[\mathbf{c}_1] & \mathrm{Cov}(\mathbf{c}_1, \mathbf{c}_2) \\ \mathrm{Cov}(\mathbf{c}_1, \mathbf{c}_2) & \mathrm{Var}[\mathbf{c}_2] \end{pmatrix}$$

The *between-classes* variance τ is:

$$\tau = \Gamma_{11} + 2 \times \Gamma_{21} \times z + \Gamma_{22} \times z^2$$

where z is the random-effects predictor (Snijders and Bosker 1999). The VPC is, therefore, a function of the random-effects predictor.
We compute the *implied* variance/covariance matrices Γ_1 and Γ_2 of random effects, relative to the two response variables:

$$\Gamma_1 = \begin{pmatrix} 0.0511 & 0.1164 \\ 0.1164 & 0.6314 \end{pmatrix} \qquad \Gamma_2 = \begin{pmatrix} 0.0702 & 0.3237 \\ 0.3237 & 1.5810 \end{pmatrix}. \tag{13}$$

Considering the variance/covariance matrices Γ_1 and Γ_2 of random effects and the variance/covariance matrix of errors $\hat{\boldsymbol{\Sigma}}$, we compute the VPCs relative to the two response variables. Figure 5 plots the VPCs as a function of the previous test score.

Of particular interest here is the way in which the ‘importance’ of the class attended increases markedly when grade 5 test scores deviate from the mean, especially for reading.

With increasing test score above the mean, especially for reading.

Considering the two marginal distributions of the class effects, we observe from Table 8 that, in the case of mathematics (first response variable), classes are divided into five subpopulations, two numerous ones containing 84.2% of the total number of classes (45.8% + 38.4%) and three smaller subpopulations containing the remaining 15% of the classes. The distribution of the class effects in reading on the four subpopulations also sees a very numerous subpopulation containing the 86.8% of the classes, followed by a subpopulation containing about the 9.5% of the classes and by two very small subpopulations containing the remaining 3.7% of the classes. By looking at the matrix $\hat{W}$ of the joint weights, we see that the joint distribution of the class effects on reading and mathematics is not uniform on the 20 mass points, but it

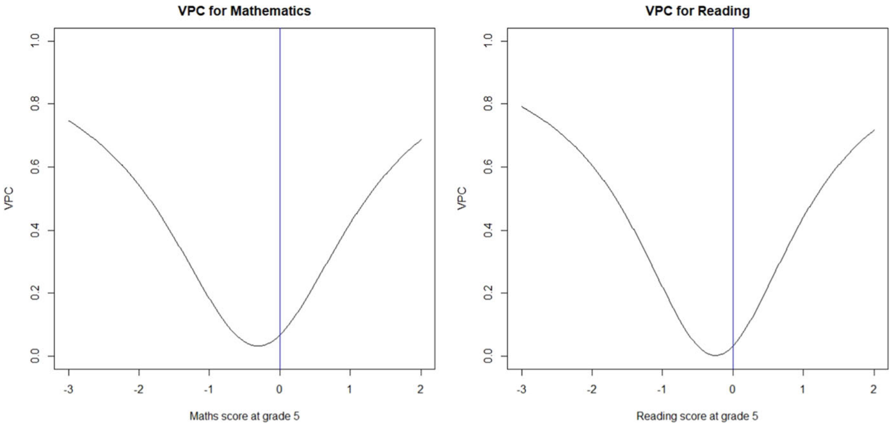


Fig. 5 VPCs relative to the two response variables (mathematics and reading, respectively), plotted as a function of the relative test score at grade 5

is mainly concentrated on certain mass points. This result further highlights the utility and the advantage of the bivariate modeling. The most numerous subpopulation is $(m = 1, k = 4)$, that contains the 45.71% of the classes, followed by subpopulation $(m = 2, k = 4)$ with the 32.20% of the classes. These two subpopulations represent the reference trend, the most common one, where classes, with respect to the other subpopulations, have the highest positive effect in reading $(k = 4)$ and a positive (but not the highest) effect in mathematics $(m = \{1, 2\})$. In terms of weights, these subpopulations are followed by subpopulation $(m = 2, k = 2)$, that contains 5.18% of the classes, that are characterized by slightly lower positive effects than the ones in the reference subpopulations. Subpopulations $(m = 3, k = 4)$ and $(m = 5, k = 4)$ contain the 2.04% and the 5.07% of the classes, respectively, and are composed by classes with the best effects both in reading and mathematics. This finding also corroborates the idea that the proportion of classes that are able to influence their students' achievement in a very positive way for both subjects is quite limited. On the opposite, subpopulations $(m = 4, k = 1)$ and $(m = 4, k = 3)$ are the worst subpopulations since students in these classes have the lowest increment in their achievements both in reading and mathematics. There are also cases where the class effects in reading and mathematics are opposite: subpopulations $(m = 5, k = 1)$ and $(m = 5, k = 3)$ are composed by classes with a very high positive effect in mathematics and a very low effect in reading; on the other side, subpopulation $(m = 4, k = 4)$ contains classes with a negative class effect in mathematics but a very high positive effect in reading.

In particular, among the entire set of different behaviors of classes, we are interested in identifying and analyzing the behaviors of the classes that significantly differ in their effects on student achievements from the ones of the reference subpopulation and, therefore, we focus our attention on four types of subpopulations:

- S_{ref} = the union of subpopulations $(m = \{1, 2\}, k = 4)$ - the reference subpopulation. It contains 843 classes, that are associated to the highest positive impact in reading and a positive impact (but not the highest) in mathematics.

Table 10 Distribution of the selected four subpopulations (S_{ref}, S_2, S_3 and S_4) in the joint distribution of the 5 × 4 subpopulations identified by the BSPEM algorithm. Except for the reference subpopulation (S_{ref}, in bold), for each subpopulation, the signs into the brackets represent the positive (+) or negative (-) class effect in mathematics and reading, respectively

	k = 1	k = 2	k = 3	k = 4
m = 1				$\boldsymbol{S_{ref}}$
m = 2				$\boldsymbol{S_{ref}}$
m = 3	$S_3(+-)$		$S_3(+-)$	
m = 4	$S_2(--)$		$S_2(--)$	$S_4(-+)$
m = 5	$S_3(+-)$		$S_3(+-)$	

- S_2 = union of subpopulations ($m = 4, k = \{1, 3\}$). It contains 16 classes, that are associated to negative impacts, with respect to the others, both in mathematics and reading.
- S_3 = union of subpopulations ($m = \{3, 5\}, k = \{1, 3\}$). It contains 13 classes, that are associated to a very positive impact in mathematics and a negative one in reading.
- S_4 = subpopulation ($m = 4, k = 4$). It contains 19 classes, that are associated to a negative impact in mathematics and a positive one in reading.

Table 10 highlights these four subpopulations in the joint distribution of the subpopulations. The subpopulations S_{ref} and S_2 contain classes that have homogeneous effects in reading and mathematics, since they exert both negative or both positive effects on their student achievements. On the other side, S_3 and S_4 contain classes that have heterogeneous effects in the two school subjects, since they exert a positive effect in mathematics and a negative one in reading and viceversa. We focus our attention on these four cases since they represent the borderline cases of all the possible interactions between class effects in mathematics and reading. Indeed, they result of great interest in the perpective of investigating eventual influences between teaching and learning dynamics in the two school subjects.

As a final remark, we must recall that in this analysis we cosider only one level of grouping, i.e. students nested within classes. As a consequence, part of the correlation that we identify among the class effects might be due to the school in which classes are nested. Future research will be dedicated to understand how schools are shaping the effectiveness of their classes in a different way.

4.3 Factors associated to the class effects

The presence of subpopulations of classes that differ in their effect on mathematics and reading student achievements might be the consequence of different class body-compositions, peers, teachers or teaching practices. These aspects may influence the class effect in reading, mathematics or both of them. Moreover, having a disadvantaged situation in one school subject learning may favor student learning in the other school subject and viceversa. Therefore, we are interested in investigating whether there are some class and teacher level variables associated to the four heterogeneous types of subpopulations. Such an exercise can be relevant for decision-makers, who can make interventions to modify schools' and classes' activities and characteristics, in search of higher levels of effectiveness. To this end, we apply a multinomial lasso logit model (Tibshirani 1996; Lokhorst 1999) by treating the class and teacher levels characteristics as covariates and the belonging of classes to the 4 subpopulations (S_{ref}, S_2, S_3, S_4) as outcome variable. This choice is driven by the fact that the number of class and teacher levels covariates is very high and we do not expect all of them to be significant. Using a lasso model allows us to select the significant covariates, addressing multicollinearity issues, and to estimate their association with the response variable. From a methodological point of view, this approach is more robust and preferable than the traditional linear modelling often used in educational research.

Denoting with Y_i the cluster of belonging of class i, for $i = 1, \ldots, N$, and considering $\mathcal{K} = \{S_{ref}, S_2, S_3, S_4\}$ the set of possible values of Y, the multinomial lasso logit model takes the following form:

$$P(Y_i = k|\mathbf{X}_i = \mathbf{x}_i) = \frac{e^{\beta_{0k}+\boldsymbol{\beta}_k^T\mathbf{x}}}{\sum_{k=1}^K e^{\beta_{0l}+\boldsymbol{\beta}_l^T\mathbf{x}_i}}, \tag{14}$$

where K is the total number of categories assumed by Y, i.e. 4, and $\mathbf{X}$ is the $N \times Q$ matrix of class and teacher levels covariates shown in Table 7. Precisely, we use as covariates all the variables presented in Table 7 except for the class body compositional variables 1^{st} `and` 2^{nd}`-gen immig`. Since we used the student-level variable `immig` as control variable in the multilevel model, using its percentage in the class as a compositional variable in the multinomial model would be misleading (Raudenbush and Willms 1995). Denoting by $\tilde{Y}$ the $N \times K$ indicator response matrix, with elements $\tilde{y}_{il} = I(y_i = l)$, the elastic-net penalized negative log-likelihood function is

$$l(\{\beta_{0k}, \beta_k\}_1^K) = -\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K}\tilde{y}_{il}(\beta_{0k} + \mathbf{x}_i^T\boldsymbol{\beta}_k) - \log(\sum_{k=1}^{K} e^{\beta_{0k}+\mathbf{x}_i^T\boldsymbol{\beta}_k})\right)\right] + \lambda\sum_{j=1}^{Q}||\boldsymbol{\beta}_j||_1, \tag{15}$$

where λ is a tuning parameter that controls the overall strength of the penalty, $\boldsymbol{\beta}$ is a $Q \times K$ matrix of coefficients, $\boldsymbol{\beta}_k$ refers to the $k-$th column (for outcome category k), and $\boldsymbol{\beta}_j$ to the $j-$th row (vector of K coefficients for variable j). We choose to perform a lasso penalty on each of the parameters.

By using cross-validation, we select the penalization term λ of the lasso regression in order to minimize the mean-squared error. The results of the lasso multinomial logit model, with the best selected choice of λ, are obtained by using the R package `glmnet` (Friedman et al. 2010) and are shown in Table 11.

According to the results of the multinomial logit model shown in Table 11, the variables that result to be significant in predicting the belonging of the classes to the four subpopulations regard contacts among teachers, the age and the gender of teachers, some aspects of the teaching methods in both mathematics and reading, the amount of hours of reading lesson and the geographical area. Classes where teachers of reading

Table 11 Results of the lasso multinomial logit regression in Eq. (14). We report in the table only the coefficients of the variables at class and teacher levels that result to be significant in the model

Variable name	S_{ref}	S_2	S_3	S_4
Teachers general questions				
Contacts among maths teachers	−0.1986			
Contacts among reading teachers	−0.1546			
Teachers personal information				
Maths teacher age	−0.0265			
Reading teacher age	−0.0046			
Reading teacher gender (male=1)				0.2316
Only for mathematics teachers				
Main teaching method 'a'	−0.1213			
Teacher written exams		−1.8238		
Only for reading teachers				
Num years of teaching here	−0.0131			
Num reading hours	−0.0188			
Summarize text	−0.0872			
Read newspapaper		0.6127		
Class information and body composition				
Area geo South	−1.6268	0.0417		

and mathematics are used to exchange views about teaching with other teachers are less likely to belong to the reference subpopulation S_{ref} (variable contacts among reading/maths teachers). The elder are the mathematics and reading teachers the less likely are classes to belong to the reference subpopulaton S_{ref} (variable maths/reading teacher age). This suggests that younger teachers are associated to worse class effects in reading and both to very positive or very negative class effects in mathematics. Classes with male reading teachers are more likely to belong to subpopulation S_4, that is the one associated to a negative impact in mathematics and a positive one in reading (variable reading teacher gender). Speaking about mathematics teaching methods, classes where teachers follow the method 'a' - teach definitions and theorems that students can apply to solve new problems - are less likely to belong to S_{ref} (the reference method is 'd' - the teacher favors the capacity of build concepts, models and theory). Classes where the mathematics teacher personally prepares the written exam for the students are less likely to belong to subpopulation S_2 (variable teacher written exam). In this case, having a mathematics teacher who does not elaborate the tests and adapt them to his/her students results to be a disadvantage, since this characteristic increases the probability of a class of being in a subpopulation with a negative effect in mathematics. Regarding the characteristics of reading, the higher is the number of hours per week dedicated to reading lesson the lower is the probability of belonging to the reference subpopulation S_{ref} (variable num reading hour). Classes where the reading teacher works in the school since many years are less likely to belong to S_{ref} (this association is in line with the one of the age of the reading teacher). Moreover, classes where the reading teacher trains students in summarinzg texts are less likely to belong to S_{ref} (variable `summarize text`). Lastly, classes where the reading teacher reads newspapers in class as part of the lesson are more likely to be associated to subpopulation S_2 (variable `read newspaper`). Classes in Southern Italy are less likely to belong to the reference subpopulation S_{ref} and are more likely to belong to S_2 (variable `area geo south`). Subpopulation S_2 contains classes with a worse effect than the ones in S_{ref} and, therefore, classes in Southern Italy have on average worse effects on student achievements than to the ones in Northern Italy.

Besides the geographical area or the number of hours of lesson per week, these results reflect the fact that personal and working characteristics of teachers are in some way associated to student learning. For instance, being a "not proactive" teacher, who simply follows the book and who does not make personalized tests, has a negative effect in mathematics and spending time in reading newspapers in class results to be a disadvantage in reading.

It is worth to remark that we do not provide any measure of uncertainty of these results since they are based on a two stage procedure and this second stage relies on the subpopulations identified at the first stage.

5 Conclusions

In this paper, we develop a bivariate semi-parametric model with random coefficients, together with an EM algorithm for estimating its parameters (BSPEM algorithm), for hierarchical data. We apply this new algorithm to Italian middle schools data of 2016/2017 for performing a classification of Italian classes. The BSPEM algorithm is the extension to the bivariate case of the SPEM algorithm presented in Masci et al. (2019). We assume the random coefficients of the model to follow a discrete distribution, where the numbers of support points of the coefficients distribution related to the multiple responses are unknown and are allowed to be different. Each group, i.e. observation at the higher level of hierarchy (classes), is assigned to one of the subpopulations identified, that characterizes the effect of the group related to the multiple response variables. The novelty and the advantage of this modeling is twofold. First, the BSPEM algorithm identifies two latent structures among the higher level of hierarchy, one related to the first response and one related to the second one (in our case, they represent test scores in two different subjects within the same class). Second, the joint modeling reveals two natures of the correlation between the two response variables: one is the correlation among the distribution of the subpopulations, that can be seen in the matrix of weights $\mathbf{w}$, that tells us how groups are distributed on the $M \times K$ mass points; the second correlation is among the unexplained variance of the two response variables, i.e. $\boldsymbol{\Sigma}_{12}$, that tells us whether in the variance of the two response variables that we are unable to explain with the model there is still correlation or not. In this perspective, the BSPEM algorithm is unique in the literature and can be applied in many classification problems, also in different fields than education, with the aim of individuating latent patterns within data or also for confirming the presence of a theoretically known number of subpopulations.

Applying the BSPEM algorithm to the achievement data of Italian middle school students, considering students as level 1 and classes as level 2, we jointly model the impact of the class on both mathematics and reading student achievements. We interpret the impact of a class as the linear relation between previous (grade 5) and current (grade 8) INVALSI test scores of students within a class, adjusting for student socio-economic index, gender and immigrant status (i.e. the value-added of class). The algorithm reveals the presence of five different trends (class effects) in mathematics and four different ones in reading. The distribution of classes on these 5×4 mass points is not uniform but it is possible to identify some more common behaviors. In particular, we distinguish classes that have a positive impacts on student achievements in both maths and reading, from the ones that have a negative one, from the ones that have heterogeneous impacts on the two school subjects.

Interested in characterizing the identified subpopulations of classes, we apply, in a second step, a lasso multinomial logit model to explain the belonging of classes to the subpopulations by means of teacher and class levels variables. It emerges that, in addition to the classical information about class body composition or peers, there are certain teacher practices or characteristics that are associated to different class impacts. In particular, the attitude, the pro-activeness and the preparation of teachers result to be effective on student learning.

The method and the results presented in this paper have three clear and important policy and managerial implications. Firstly, it is useful to classify the classes in groups on the basis of their likely effect on student achievement, instead of creating "rankings" among them. This way, the characteristics of groups can be analysed, and decision makers can have clear indications about how to intervene to try boosting the effectiveness of educational activities. For example, our results point at demonstrating that classes where the effects on achievement are more positive are those in which teachers adopt a more proactive in building concepts, methods and theories. Secondly, the effectiveness of classes must be judged on the basis of their joint effect on different subjects, in a multidimensional perspective. Our results indicate that many classes are able to exert a positive effect on students' achievement in one subject but not the other. The proportion of classes that contribute very positively to achievement in both reading and mathematics is quite limited (around 10%), and they should serve as a benchmark and reference point to understand the key features that make them particularly effective. Anyway, most of previous literature in the field focuses on one subject at a time, so neglecting a lot of the complex interaction in teaching and educational practices that have an effect on students' results - and our work overcomes this problem. Thirdly, background individual characteristics of the students are confirmed to be very important in influencing their academic results. The estimate of classes' effects that we provide are determined net of students' characteristics, but a necessary development of our methodology will be to study more profoundly the interaction between individual features' and classes' characteristics and activities. This way, the proposed method could provide useful insights to understand which are the likely results of moving students between classes.

A limitation of our study, determined by data availability, is that we do not have information about multiple classes within the same school. Potentially, the proposed model can be extended to the case of a three-level hierarchy: students as level 1, classes as level 2 and schools as level 3. In so doing, thanks to the random effect at school level, the effect of the second level, i.e. the class effect, would be the within-school class effect, allowing to model both between schools and within-school variabilities. We believe in the potential of this approach but, since the available INVALSI data that we presented in Sect. 2 regard classes that are all nested within different schools (each class corresponds to a different school), we could not consider three different levels in our application. For this reason, an interesting development of our research effort will consist in obtaining new data and exploring how the information about the clustering in different schools influences the heterogeneity of classes' effectiveness, adjusting for individual students' characteristics.

Summing up, the present study pares the way for extensions towards better understanding of the educational production process, in particular for modelling heterogeneity of effects within classes and schools.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Appendix A

The EM algorithm for bivariate semi-parametric linear models with random coefficients

The EM algorithm that we propose to estimate the parameters of the model in (3) is the generalization for the bivariate case of the one proposed in Masci et al. (2019). It alternates two steps: the expectation step (E step) in which we compute the conditional expectation of the likelihood function with respect to the random coefficients, 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 Eq. (4) and it continues until the convergence. The update of the parameters is the following:

$$w_{mk}^{(up)} = \frac{1}{N}\sum_{i=1}^{N} W_{imk} \qquad \text{for } m = 1, \ldots, M, \quad k = 1, \ldots, K \tag{16}$$

and

$$(\boldsymbol{B}^{(up)}, \mathbf{C}_{mk}^{(up)}, \boldsymbol{\Sigma}^{(up)}) = \underset{\mathbf{B}, \mathbf{C}_{mk}, \boldsymbol{\Sigma}}{\arg\max} \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{i=1}^{N} W_{imk} \ln p(\mathbf{y}_i | \boldsymbol{B}, \boldsymbol{\Sigma}, \mathbf{C}_{mk}) \tag{17}$$

where

$$W_{imk} = \frac{w_{mk}\, p(\mathbf{y}_i | \boldsymbol{B}, \boldsymbol{\Sigma}, \mathbf{C}_{mk})}{\sum_{m=1}^{M} \sum_{k=1}^{K} w_{mk}\, p(\mathbf{y}_i | \boldsymbol{B}, \boldsymbol{\Sigma}, \mathbf{C}_{mk})} \tag{18}$$

and

$$\begin{aligned} p(\mathbf{y}_i | \mathbf{B}, \boldsymbol{\Sigma}, \mathbf{C}_{mk}) &= \frac{1}{\sqrt{|det(2\pi\boldsymbol{\Sigma})|^{n_i}}} \\ &\times \exp\left\{ \sum_{j=1}^{n_i} -\frac{1}{2} \begin{pmatrix} y_{1,ij} - c_{1,1m} - \sum_{p=1}^{P} \beta_{1p} x_{1p,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \sum_{p=1}^{P} \beta_{2p} x_{2p,ij} - c_{2,2k} z_{2,ij} \end{pmatrix}^{T} \boldsymbol{\Sigma}^{-1} \right. \\ &\times \left. \begin{pmatrix} y_{1,ij} - c_{1,1m} - \sum_{p=1}^{P} \beta_{1p} x_{1p,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \sum_{p=1}^{P} \beta_{2p} x_{2p,ij} - c_{2,2k} z_{2,ij} \end{pmatrix} \right\}. \end{aligned} \tag{19}$$

For each response variable, W represents the the conditional weight matrix: the coefficient W_{imk} represents the probability of $\mathbf{1}_i$ being equal to $\mathbf{C}_{mk}$ conditionally to observations $\mathbf{y}_i$ and given the fixed coefficient $\boldsymbol{B}$ and the variance/covariance matrix $\boldsymbol{\Sigma}$. Indeed, since $w_{mk} = p(\mathbf{1}_i = \mathbf{C}_{mk})$, then

$$\begin{aligned} W_{imk} &= \frac{w_{mk}\, p(\mathbf{y}_i | \boldsymbol{B}, \boldsymbol{\Sigma}, \mathbf{C}_{mk})}{\sum_{m=1}^{M} \sum_{k=1}^{K} w_{mk}\, p(\mathbf{y}_i | \boldsymbol{B}, \boldsymbol{\Sigma}, \mathbf{C}_{mk})} = \frac{p(\mathbf{1}_i = \mathbf{C}_{mk})\, p(\mathbf{y}_i | \boldsymbol{B}, \boldsymbol{\Sigma}, \mathbf{C}_{mk})}{p(\mathbf{y}_i | \boldsymbol{B}, \boldsymbol{\Sigma})} \\ &= \frac{p(\mathbf{y}_i, \mathbf{1}_i = \mathbf{C}_{mk} | \boldsymbol{B}, \boldsymbol{\Sigma})}{p(\mathbf{y}_i | \boldsymbol{B}, \boldsymbol{\Sigma})} = p(\mathbf{1}_i = \mathbf{C}_{mk} | \mathbf{y}_i, \boldsymbol{B}, \boldsymbol{\Sigma}). \end{aligned} \tag{20}$$

Therefore, in order to compute the point $\mathbf{C}_{mk}$ for each group i, for $i = 1, \dots, N$, we maximize the conditional probability of $\mathbf{1}_i$ given the observations $\mathbf{y}_i$, the coefficient $\boldsymbol{B}$ and the error variance/covariance matrix $\boldsymbol{\Sigma}$. So that, the estimation of the coefficients $\mathbf{1}_i$ for each group i is obtained maximizing W_{imk} over m and k, that is

$$\hat{\mathbf{1}}_i = \mathbf{C}_{\tilde{mk}} \quad \text{where} \quad \tilde{mk} = \underset{m,k}{\arg\max}\, W_{imk} \quad i = 1, \dots, N. \tag{21}$$

The maximization in Eq. (17) involves two steps and it is done iteratively. In the first step, we compute the *arg-max* with respect to the support points $\mathbf{C}_{mk}$, keeping $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ fixed to the last computed values. In this way, we can maximize the expected log-likelihood with respect to all support points $\mathbf{C}_{mk}$ separately, that means

$$\mathbf{C}_{mk}^{(up)} = \underset{\mathbf{C}}{\arg\max} \sum_{i=1}^{N} W_{imk} \ln p(\mathbf{y}_i | \boldsymbol{B}, \boldsymbol{\Sigma}, \mathbf{C}_{mk}) \qquad m = 1, \dots, M \quad k = 1, \dots, K. \tag{22}$$

Since we are considering the linear case, the maximization step is done in closed-form.[13] In the second step, we fix the support points of the random coefficients distribution computed in the previous step and we compute the *arg-max* in Eq. (17) with respect to $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$. Again, this step is done in closed-form.

The initialization of the support points of the discrete distribution S^* and the criteria for the convergence of the EM algorithm are the direct extension of the ones chosen in Masci et al. (2019) for the bivariate case. In particular, the algorithm starts considering N support points for the random coefficients and a starting estimate for the fixed coefficients, for both the response variables. These parameters are chosen in the following way:

- random coefficients: for each response variable, the starting N support points are obtained fitting a simple linear 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 $N \times N$ support points;
- fixed coefficients: the starting values of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are estimated by fitting a unique bivariate linear regression on the entire population (i.e. without considering the nesting of the observations within 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 coefficients distribution is done in the following way:

- we choose a number $N^* < N$ of support points, that is the same for both the two response variables;
- for each response variable, we extract N^* points from a uniform distribution with support on the entire range of possible values for each parameter, that is estimated by fitting N distinct linear regressions for each one of the N groups, as before, and identifying the minimum and the maximum values;
- we uniformly distribute the weights on these $N^* \times N^*$ support points.

The $M \times K$ matrix of weights, that is composed by the elements w_{mk} previously described, represents the joint distribution of groups across the bivariate clusters and, by summing over rows and columns respectively, it represents the marginal distribution of the groups across the univariate clusters, for each single response variable.

During the iterations, the EM algorithm performs the support reduction of the discrete distribution of random coefficients, in order to identify $M \times K$ mass points (starting from $N \times N$ mass points), where both M and K are smaller than N. The support reduction is made standing on two criteria. The former is that we fix a threshold value D and if two mass points are closer, in terms of euclidean distance, than D, they collapse to a unique point. This procedure is separately applied to the clusters related to the first and second response variable respectively. In particular, considering, for example, the case of the first response variable, if two mass points $\mathbf{c}_{1,h}$ and $\mathbf{c}_{1,g}$, for $h, g = 1, \ldots, M$, are closer than D, they collapse to a unique point $\mathbf{c}_{1,(hg)}$, where

[13] Closed-form calculations of model parameters can be found in Masci et al. (2019).

$\mathbf{c}_{1,(hg)} = \frac{\mathbf{c}_{1,h}+\mathbf{c}_{1,g}}{2}$. Consequently, $M^{new} = M^{old} - 1$, the new marginal weight is obtained as $w_{1,(hg)} = w_{1,h} + w_{1,g}$ and the joint weights $w_{(hg)k} = w_{hk} + w_{gk}$, for $k = 1, \dots, K$. The same criterion applies to the clusters related to the second response variable. The first two masses collapsing 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 so forth. note that, even if D_1 and D_2 assume the same value for the clusters related to the two response variables, the procedure might lead to different number of mass points M and K. Regarding this procedure, it is also worth to notice that it is important to standardize the covariates, in order to make the computation based on the Euclidean distance fair. Indeed, in order for the metric to be consistent, units of measurement of the coefficients have to be the same. The second criterium is that, starting from a given iteration up to the end, we fix a threshold value $\tilde{w}$ and we remove mass points with marginal weights $w_{1,m} \leq \tilde{w}$, for $m = 1, \dots, M$ and $w_{2,k} \leq \tilde{w}$, for $k = 1, \dots, K$ or that are not associated to any subpopulation. D and $\tilde{w}$ are two parameters that tune the estimates of the subpopulations. The choice of these two tuning parameters is not trivial, but it can be driven by the application aim. Setting a minimum weight $\tilde{w} > 0$ serves to delete those groups that have anomalous behaviours, different from the identified main subpopulations. Researchers might set a $\tilde{w} > 0$ when their interest is in the identificaton of big subpopulations and not in the outlier groups (that can therefore be deleted) or when the number of groups is prohibitively high to consider all the singleton groups. On the opposite, setting $\tilde{w} = 0$ allows to take into consideration all types of subpopulations, included the ones composed by single groups (interpreted as outlier groups). As introduced in Sect. 2, the value D sets the minimum heterogeneity across subpopulations and can be chosen depending on how much we want to be sensitive on the differences across groups. The choice of D can also be supported by evaluating the uncertainty of classification (with which the algorithm classifies groups into subpopulations), measured by the entropy of the rows of the 3-dimensional array W. W is the $(N \times M \times K)-$dimensional array of conditional joint weights. In the best case, i.e. when the algorithm assigns each group i to a joint subpopulation mk with probability 1, each row i, for $i = 1, \dots, N$, of the array W would be composed of $(M \times K - 1)$ values equal to 0 and a value equal to 1. In this scenario, the entropy $E_i = -\sum_{m=1}^{M} \sum_{k=1}^{K} W_{imk} \ln(W_{imk})$ of each row i of the array W would be equal to 0. The more the distribution of the weights is uniform on the $M \times K$ mass points, the higher is the entropy and, therefore, the higher is the uncertainty of classification. The worst case happens when the distribution of the weights of a group i is uniform on the $M \times K$ subpopulations ($W_{imk} = \frac{1}{M \times K}$), which corresponds to an entropy $E_i = -\sum_{m=1}^{M} \sum_{k=1}^{K} \frac{1}{M \times K} \ln(\frac{1}{M \times K}) = -\ln(\frac{1}{M \times K})$. The entropy of the matrix W constitutes a driver for the choice of the tuning parameter D, suggesting a lower bound for D that minimizes the entropy. Further insights on the choice of these parameters can be found in Masci et al. (2019).

The sketch of the BSPEM algorithm is shown in Algorithm 1. At each iteration a, the algorithm, given the estimated number of mass points, estimates all the parameters in Eq. (3) in an iterative way, updating both fixed and random coefficients, until convergence or until it reaches the maximum number of sub-iterations fixed

Algorithm 1: EM algorithm for bivariate semi-parametric models with random coefficients

```
input  : Initial estimates for $(\mathbf{C}_{11}^{(0)}, \dots, \mathbf{C}_{MK}^{(0)})$ and $(w_{11}^{(0)}, \dots, w_{MK}^{(0)})$, with $M = N$ and $K = N$;
         Initial estimates for $\boldsymbol{B}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$;
         Tolerance parameters $D_1$, $D_2$, $\tilde{w}_1$, $\tilde{w}_2$, tollR, tollF, it, it1, itmax.
output: Final estimates of $\mathbf{C}_{mk}^{(a)}$, $w_{mk}^{(a)}$, for $m = 1, \dots, M$, $k = 1, \dots, K$, $\boldsymbol{B}^{(a)}$ and $\boldsymbol{\Sigma}^{(a)}$.
a=1; conv1=0; conv2=0;
while $(conv1 == 0$ or $conv2 == 0$ & $a <$ it$)$ do
    compute the distance matrices DIST1 and DIST2 for both the subpopulations distribution
    (where, e.g.,for the first response variable, $DIST1_{st} = \sqrt{(c_{1,1s} - c_{1,1t})^2 + (c_{1,2s} - c_{1,2t})^2}$ is
    the euclidean distance between each couple of mass points $s, t$ $\forall s, t = 1, \dots, M, s \neq t$);
    if $(DIST1_{st} < D_1$ & $DIST1_{st} = min(DIST1)$   $(\forall s, t = 1, \dots, M, s \neq t))$ then
        collapse marginal masses $s$ and $t$ to a unique mass point;
    if $(DIST2_{st} < D_2$ & $DIST2_{st} = min(DIST2)$   $(\forall s, t = 1, \dots, K, s \neq t))$ then
        collapse marginal masses $s$ and $t$ to a unique mass point;
    compute the new distance matrices DIST1 and DIST2;
    if $conv1 == 1$ or $a \geq$ it1 then
        if $w_{1,m}^{(a)} \leq \tilde{w}_1$   $(\forall m = 1, \dots, M)$ then
            delete marginal mass point $m$;
            reparameterize the weights;
        if $w_{2,k}^{(a)} \leq \tilde{w}_2$   $(\forall k = 1, \dots, K)$ then
            delete marginal mass point $k$;
            reparameterize the weights;
        if no changes are done then
            $conv2 = 1$;
    given $\mathbf{C}_{mk}^{(a-1)}$, $w_{mk}^{(a-1)}$ for $m = 1, \dots, M$ and $k = 1, \dots, K$, $\boldsymbol{B}^{(a-1)}$ and $\boldsymbol{\Sigma}^{(a-1)}$, compute the
    matrix W according to Eq. (20);
    update the weights $w_{11}^{(a)}, \dots, w_{MK}^{(a)}$ according to Eq. (16);
    $\boldsymbol{B}^{(a,0)} = \boldsymbol{B}^{(a-1)}$;
    $\boldsymbol{\Sigma}^{(a,0)} = \boldsymbol{\Sigma}^{(a-1)}$;
    $\mathbf{C}_{mk}^{(a,0)} = \mathbf{C}_{mk}^{(a-1)}$;
    $w_{mk}^{(a,0)} = w_{mk}^{(a-1)}$;
    keeping $\boldsymbol{B}^{(a,0)}$ and $\boldsymbol{\Sigma}^{(k,0)}$ fixed, update the $M \times K$ support points $\mathbf{C}_{11}^{(a,1)}, \dots, \mathbf{C}_{MK}^{(a,1)}$ according
    to Eq. (17);
    keeping $\mathbf{C}_{mk}^{(a,1)}$, $w_{mk}^{(a,0)}$ for $m = 1, \dots, M$ and $k = 1, \dots, K$ fixed, update $\boldsymbol{B}^{(a,1)}$ and $\boldsymbol{\Sigma}^{(a,1)}$
    according to Eq. (17);
    j=1;
    while $(|\boldsymbol{B}^{(a,j-1)} - \boldsymbol{B}^{(a,j)}| \geq tollF$  or  $|\boldsymbol{\Sigma}^{(a,j-1)} - \boldsymbol{\Sigma}^{(a,j)}| \geq$
    $tollF$  or  $|\mathbf{C}_{mk}^{(a,j-1)} - \mathbf{C}_{mk}^{(a,j)}| \geq tollR)$  &  $j \leq$ itmax do
        j=j+1;
        keeping $\boldsymbol{B}^{(a,j-1)}$ and $\boldsymbol{\Sigma}^{(a,j-1)}$ fixed, update the $M \times K$ support points $\mathbf{C}_{11}^{(a,j)}, \dots, \mathbf{C}_{MK}^{(a,j)}$
        according to Eq. (17);
        keeping $\mathbf{C}_{mk}^{(a,j)}$, $w_{mk}^{(a,j-1)}$ for $m = 1, \dots, M$ and $k = 1, \dots, K$ fixed, update $\boldsymbol{B}^{(a,j)}$ and
        $\boldsymbol{\Sigma}^{(a,j)}$ according to Eq. (17);
    set $\mathbf{C}_{mk}^{(a)} = \mathbf{C}_{mk}^{(a,j)}$ for $m = 1, \dots, M$ and $k = 1, \dots, K$, $\boldsymbol{B}^{(a)} = \boldsymbol{B}^{(a,j)}$, $\boldsymbol{\Sigma}^{(a)} = \boldsymbol{\Sigma}^{(a,j)}$;
    estimate subpopulation $mk$ for each group $i$ according to Eq. (21);
    if $(\boldsymbol{B}^{(a)} - \boldsymbol{B}^{(a-1)} < tollF)$ & $(\boldsymbol{\Sigma}^{(k)} - \boldsymbol{\Sigma}^{(k-1)} < tollF)$ & $(\mathbf{C}_{mk}^{(a)} - \mathbf{C}_{mk}^{(a-1)} < tollR)$ then
        $conv1 = 1$;
    a= a+1;
```

a priori for this stage (`itmax`). At the beginning of the iterative process, the algorithm performs the dimensional reduction of the mass points standing only on 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 standing also on the criterion of the minimum weight $\tilde{w}$. The final convergence is reached when 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 `it`. In particular, we fix the tolerance values for the estimates of both the fixed and random coefficients to `tollF` and `tollR` respectively, which depend on the scale of the parameters. The usage of the maximum number of iterations `it`, `it1` and `itmax` is merely to avoid an infinite loop and their values depend on the complexity of the data and on the consequent convergence rate. The code is implemented using the R software (R Core Team 2014) and it is available upon request.

Appendix B

Further simulations: BSPEM algorithm convergence and robustness with respect to group sizes

In this section, we repeat the first simulation study presented in Sect. 3, but considering $n_i = 20$ instead of $n_i = 100$, for $i = 1, \dots, 100$. We add this simulation in order to check whether the method converges and identifies the simulated parameters also with a smaller number of observations within groups.[14] Table 12 reports the parameters estimated by the BSPEM algorithm, obtained as the average over the 100 runs. On average, the algorithm converges in 7 iterations and it always identifies the correct number of clusters for both the two response variables.

The estimated weights matrix and the variance/covariance matrix $\boldsymbol{\Sigma}$ with its Mean Squared Error are the following:

$$\hat{W} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix} \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.9939 & 0.0049 \\ 0.0049 & 0.9919 \end{pmatrix}$$

$$MSE_{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0009 & 0.0005 \\ 0.0005 & 0.0007 \end{pmatrix}.$$

Results are consistent with the ones obtained by considering $n_i = 100$ observations within each group, shown in Table 2. Table 12 shows that the estimated parameters are very close to the real ones used to simulate the data and their MSE is also very low.

[14] We choose $n_i = 20$ to simulate the average group size in the case study.

Table 12 Values of the parameters of Eq. (5) estimated by the BSPEM algorithm, obtained as the average over the 100 runs (for each parameter we also report its Mean Square Error in brackets) and considering $n_i = 20$

	First response parameters	Second response parameters
$i = 1, \dots, 33$	$\hat{c}_{1,11} = 4.99334$	$\hat{c}_{2,11} = 3.000301$
	$(MSE_{1,11} = 0.00181)$	$(MSE_{2,11} = 0.00103)$
	$\hat{c}_{1,21} = 9.99898$	$\hat{c}_{2,21} = 0.99858$
	$(MSE_{1,21} = 0.00171)$	$(MSE_{2,21} = 0.00086)$
	$\hat{\beta}_1 = 3.00171$	$\hat{\beta}_2 = 1.99854$
	$(MSE_{\beta_1} = 0.00073)$	$(MSE_{\beta_2} = 0.00065)$
$i = 34, \dots, 66$	$\hat{c}_{1,12} = 1.99934$	$\hat{c}_{2,11} = 3.00030$
	$(MSE_{1,12} = 0.00167)$	$(MSE_{2,11} = 0.00103)$
	$\hat{c}_{1,22} = 4.99161$	$\hat{c}_{2,21} = 0.99858$
	$(MSE_{1,22} = 0.00229)$	$(MSE_{2,21} = 0.00086)$
	$\hat{\beta}_1 = 3.00171$	$\hat{\beta}_2 = 1.99854$
	$(MSE_{\beta_1} = 0.00073)$	$(MSE_{\beta_2} = 0.00065)$
$i = 67, \dots, 100$	$\hat{c}_{1,13} = -0.00168$	$\hat{c}_{2,12} = -0.00470$
	$(MSE_{1,13} = 0.00147)$	$(MSE_{2,12} = 0.00171)$
	$\hat{c}_{1,23} = -1.99569$	$\hat{c}_{2,22} = -3.00502$
	$(MSE_{1,23} = 0.00246)$	$(MSE_{2,22} = 0.00137)$
	$\hat{\beta}_1 = 3.00171$	$\hat{\beta}_2 = 1.99854$
	$(MSE_{\beta_1} = 0.00073)$	$(MSE_{\beta_2} = 0.00065)$

Colors represent the different subpopulations identified by the algorithm. The algorithm identifies three subpopulations (M=3) for the first response and two subpopulations for the second one (K=2)

Comparing the MSEs of Tables 2 and 12, emerges that the MSEs in Table 2 usually differ from 0 at the fourth decimal digit while the MSEs in Table 12 usually differ from 0 at the third one. The BSPEM method results to be robust with respect to group sizes.

References

Agasisti T, Ieva F, Paganoni AM (2017) Heterogeneity, school-effects and the north/south achievement gap in italian secondary education: evidence from a three-level mixed model. Stat Methods Appl 26(1):157–180

Agasisti T, Vittadini G (2012) Regional economic disparities as determinants of student's achievement in italy. Res Appl Econ 4(2):33

Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. Stat Comput 6(3):251–262

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

Barndorff-Nielsen O (1965) Identifiability of mixtures of exponential families. J Math Anal Appl 12(1):115–121

Belfi B, Goos M, De Fraine B, Van Damme J (2012) The effect of class composition by gender and ability on secondary school students' school well-being and academic self-concept: A literature review. Educ Res Rev 7(1):62–74

Bietenbeck J (2014) Teaching practices and cognitive skills. Labour Econ 30:143–153

Brewer DJ, Goldhaber D (1997) Why don't schools and teachers seem to matter?: Assessing the impact of unobservables on educational productivity. J Hum Res Summer

Cramér H (1999) Mathematical methods of statistics (43). Princeton University Press, New Jersey

Dahl DB (2006) Model-based clustering for expression data via a dirichlet process mixture model. Bayesian Inference for Gene Expression and Proteomics 4:201–218

Dar Y, Resh N (1986) Classroom intellectual composition and academic achievement. Am Educ Res J 23(3):357–374

Dar Y, Resh N (2018) Classroom composition and pupil achievement (1986): A study of the effect of ability-based classes. Routledge

David R, Teddlie C, Reynolds D (2000) The international handbook of school effectiveness research. Psychology Press

De Witte K, Van Klaveren C (2014) How are teachers teaching? a nonparametric approach. Educ Econ 22(1):3–23

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw **33**(1): 1 www.jstatsoft.org/v33/i01/

Goldhaber DD, Brewer DJ (1997) Why don't schools and teachers seem to matter? assessing the impact of unobservables on educational productivity. J Hum Resour pp 505–523

Goldstein H, Browne W, Rasbash J (2002) Partitioning variation in multilevel models. Understanding statistics: statistical issues in psychology, education, and the social sciences 1(4):223–231

Grilli L, Pennoni F, Rampichini C, Romeo I (2016) Exploiting timss and pirls combined data: multivariate multilevel modelling of student achievement. Ann Appl Stat 10(4):2405–2426

Grilli L, Rampichini C (2009) Multilevel models for the evaluation of educational institutions: a review. Statistical methods for the evaluation of educational services and quality of products. Springer, Berlin, pp 61–80

Hanushek EA (1992) The trade-off between child quantity and quality. J Polit Econ 100(1):84–117

Leckie G (2018) Avoiding bias when estimating the consistency and stability of value-added school effects. J Educ Behav Stat 43(4):440–468

Leckie G, Goldstein H (2017) The evolution of school league tables in England 1992–2016:contextual value-added, expected progress and progress 8. Brit Educ Res J 43(2):193–212

Lin LI (2000) Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Stat Med 19(2):255–270

Lindsay BG (1983a) The geometry of mixture likelihoods: a general theory. Ann Stat 11(1):86–94

Lindsay BG (1983b) The geometry of mixture likelihoods, part ii: the exponential family. Ann Stat 11(3):783–792

Lokhorst J (1999) The lasso and generalised linear models. The University of Adelaide, Australia, Honors Project

Martineau JA (2006) Distorting value added: the use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. J Educ Behav Stat 31(1):35–62

Masci C, Ieva F, Agasisti T, Paganoni AM (2016) Does class matter more than school? evidence from a multilevel statistical analysis on italian junior secondary school students. Socio-Econ Plan Sci 54:47–57

Masci C, Ieva F, Agasisti T, Paganoni AM (2017) Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. J Appl Stat 44(7):1296–1317

Masci C, Paganoni AM, Ieva F (2019) Semiparametric mixed effects models for unsupervised classification of italian schools. J R Stat Soc Ser A Stat Soc 182(4):1313–1342

McCaffrey DF, Lockwood J, Koretz D, Louis TA, Hamilton L (2004) Models for value-added modeling of teacher effects. J Educ Behav Stat 29(1):67–101

McCulloch C, Lin H, Slate E, Turnbull B (2002) Discovering subpopulation structure with latent class mixed models. Stat Med 21(3):417–429

Meyer RH (1997) Value-added indicators of school performance: a primer. Econ Educ Rev 16(3):283–301

Muthén B (2004) Latent variable analysis. Sage Handb Quant Methodol Soc Sci 345:368

Muthén B, Asparouhov T (2015) Growth mixture modeling with non-normal distributions. Stat Med 34(6):1041–1058

Muthén B, Shedden K (1999) Finite mixture modeling with mixture outcomes using the em algorithm. Biometrics 55(2):463–469

Nagin DS (1999) Analyzing developmental trajectories: a semiparametric, group-based approach. Psychol Methods 4(2):139

Palardy GJ (2008) Differential school effects among low, middle, and high social class composition schools: A multiple group, multilevel latent growth curve analysis. Sch Eff Sch Improv 19(1):21–49

Parsons E, Koedel C, Tan L (2018) Accounting for student disadvantage in value-added models. J Educ Behav Stat 4:144–179

Perry T (2016) English value-added measures: examining the limitations of school performance measurement. Brit Educ Res J 42(6):1056–1080

Pinheiro JC, Bates DM (2000) Linear mixed-effects models: basic concepts and examples. Mixed-effects models in S and S-Plus 3–56

Puccetti G (2019) Measuring linear correlation between random vectors. Available at SSRN 3116066

R Core Team (2014) R: A language and environment for statistical computing. Vienna, Austria . http://www.R-project.org/

Raudenbush SW, Willms J (1995) The estimation of school effects. J Educ Behav Stat 20(4):307–335

Rights JD, Sterba SK (2016) The relationship between multilevel models and non-parametric multilevel mixture models: Discrete approximation of intraclass correlation, random coefficient distributions, and residual heteroscedasticity. Br J Math Stat Psychol 69(3):316–343

Rivkin SG, Hanushek EA, Kain JF (2005) Teachers, schools, and academic achievement. Econometrica 73(2):417–458

Rockoff JE (2004) The impact of individual teachers on student achievement: Evidence from panel data. Am Econ Rev 94(2):247–252

Sani C, Grilli L (2011) Differential variability of test scores among schools: multilevel analysis of the fifth-grade invalsi test using heteroscedastic random effects. J Appl Quant Methods 6(4):88–99

Schagen I, Schagen S (2005) Combining multilevel analysis with national value-added data sets - a case study to explore the effects of school diversity. Brit Educ Res J 31(3):309–328

Schwerdt G, Wuppermann AC (2011) Is traditional teaching really all that bad? a within-student between-subject approach. Econ Educ Rev 30(2):365–379

Snijders TA, Bosker RJ (1999) Multilevel analysis: an introduction to basic and advanced multilevel modeling. sage

Strand S (1997) Pupil progress during key stage 1: a value added analysis of school effects. Brit Educ Res J 23(4):471–487

Teicher H (1963) Identifiability of finite mixtures. Ann Math Stat pp 1265–1269

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Stat Soc Ser B Methodol pp 267–288

Timmermans AC, Bosker RJ, de Wolf IF, Doolaard S, van der Werf MP (2014) Value added based on educational positions in dutch secondary education. Brit Educ Res J 40(6):1057–1082

Vermunt JK, Magidson J (2002) Latent class cluster analysis. Appl Latent Class Anal 11:89–106

Wayne AJ, Youngs P (2003) Teacher characteristics and student achievement gains: a review. Rev Educ Res 73(1):89–122

Wenglinsky H (2002) The link between teacher classroom practices and student academic performance. Educ Policy Anal Arch 10:12

Winkler DR (1975) Educational achievement and school peer group composition. J Hum Resour pp 189–204

Yang M, Goldstein H, Browne W, Woodhouse G (2002) Multivariate multilevel analyses of examination results. J R Stat Soc Ser A Stat Soc 165(1):137–153