MOX-Report No. 24/2019

Evaluating class and school effects on the joint achievements in different subjects: a bivariate semi-parametric mixed-effects model

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

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

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

Evaluating class and school effects on the joint achievements in different subjects: a bivariate semi-parametric mixed-effects model

Masci C.[♮], Ieva F.[♮], Agasisti T.[♢], Paganoni A. M.[♮]

June 6, 2019

[♮] MOX - Modelling and Scientific Computing, Department of Mathematics,
Politecnico di Milano, via Bonardi 9, Milano, Italy
chiara.masci@polimi.it
francesca.ieva@polimi.it
anna.paganoni@polimi.it
[♢] DIG - Department of Management, Economics and Industrial Engineering,
Politecnico di Milano, via Lambruschini 4/b, Milano, Italy
tommaso.agasisti@polimi.it

Abstract

This paper proposes an innovative statistical method to measure the impact of the class/school on its student achievements in multiple subjects. We propose a semi-parametric mixed-effects model with a bivariate response variable, where the random effects are assumed to follow a discrete distribution with an unknown number of support points, together with an Expectation-Maximization algorithm to estimate its parameters. The bivariate setting allows to estimate the distributions of the model coefficients related to each response variable as well as their joint distribution. 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 two different subjects (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.

Key-words: Semi-parametric mixed-effects model; EM algorithm; Clustering; LASSO regression; School and class effects; Student achievements; Teaching practices.

Acknowledgements

The authors are grateful to the helpful comments received at the *International Conference on Education Economics (Budapest 2018)* and at the *International Society of Nonparametric Statistics Conference (Salerno, 2018)*.

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 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 peformance of schools populated by students with a comparable level of prior achievement (and other student characteristics) (Raudenbush & Willms, 1995; Schagen & Schagen, 2005; Timmermans, Bosker, de Wolf, Doolaard, & van der Werf, 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, Lockwood, Koretz, Louis, & Hamilton, 2004; Martineau, 2006). The most recent literature about this topic (Perry, 2016; Leckie & Goldstein, 2017; Parsons, Koedel, & Tan, 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, Hanushek, & Kain, 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 & Brewer, 1997; Wenglinsky, 2002; Schwerdt & Wuppermann, 2011; De Witte & 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, Ieva, & Paganoni, 2017; Grilli & Rampichini, 2009; Masci, Ieva, Agasisti, & Paganoni, 2016, 2017; Sani & 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, driven by certain school characteristics (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 *school effect* in the context of within-school heterogeneity. A similar problem is discussed in (Masci, Paganoni, & Ieva, 2019), where the authors apply a

multilevel linear model, but, instead of following a classical parametric approach, they follow a semi-parametric approach in which they develop a semi-parametric mixed-effects (two-level, where students are nested within schools) model able to identify a latent structure among the higher 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, extending the new statistical model presented in (Masci et al., 2019), we propose a study that is innovative both from a methodological and an interpretative point of view. We develop the bivariate version of the Expectation-Maximization algorithm for semi-parametric mixed-effects models (SPEM algorithm) presented in (Masci et al., 2019), i.e. we extend it 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 coefficients of random effects, 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 mixed-effects linear model. The distribution of the coefficients of random effects 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 effects related to each one of the two response variables and, moreover, to estimate the joint distribution of random effects related to the two response variables, investigating the correlation among them. Read in the context of the educational literature on school value-added, it means that 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). 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 methodology proposed here is completely 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 & Shedden, 1999; Nagin, 1999) and of Latent Class Mixture Models (LCMM) (McCulloch, Lin, Slate, & Turnbull, 2002; Vermunt & 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 et al., 2000; Muthén & 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.

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) even though it is formally presented in its complete three levels form (students, classes and schools). 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 effects (class effects) for both the first and the second response. Moreover, by looking at the joint distribution of these random effects, 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 Section 2, we present the bivariate semi-parametric two-level linear model. In Section 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 BSPEM algorithm to it and showing its results and (iii) analyzing a posteriori the characteristics of the identified subpopulations of classes. In Section 5, we draw policy implications and conclusions.

2 Model and methods: the bivariate semi-parametric mixed-effects linear model

In this section, we present the bivariate semi-parametric mixed-effects linear model[1].

Consider a bivariate three-level linear model (considering only the random intercept at the highest level of hierarchy), where each bivariate observation j, for $j = 1, \dots, n_{il}$, is nested within a group i, for $i = 1, \dots, I_l$, that is in turn nested within a group l, for $l = 1, \dots, L$. The model takes the following form:

$$\begin{aligned} \mathbf{Y}_{il} = \begin{pmatrix} \mathbf{y}_{1,il} & \mathbf{y}_{2,il} \end{pmatrix} = \mathbf{X}_{il} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}^T + \mathbf{Z}_{il} \begin{pmatrix} \mathbf{c}_{1,il} \\ \mathbf{c}_{2,il} \end{pmatrix}^T + \boldsymbol{\alpha}_l + \boldsymbol{\epsilon}_{il} \\ i = 1, \dots, I_l, \quad l = 1, \dots, L \\ \boldsymbol{\alpha}_l^T = \begin{pmatrix} \boldsymbol{\alpha}_{1,l} \\ \boldsymbol{\alpha}_{2,l} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_\alpha) \qquad \boldsymbol{\epsilon}_{il}^T = \begin{pmatrix} \boldsymbol{\epsilon}_{1,il} \\ \boldsymbol{\epsilon}_{2,il} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind. \end{aligned} \tag{1}$$

where $N = \sum_{l=1}^{L} I_l$ is the total number of level 2 groups and $J = \sum_{l=1}^{L} \sum_{i=1}^{I_l} n_{il}$ is the total number of bivariate observations[2]. The components of model (1) are the following:

- $\mathbf{Y}_i = \begin{pmatrix} y_{1,1il}, \dots, y_{1,n_{il}il} \\ y_{2,1il}, \dots, y_{2,n_{il}il} \end{pmatrix}^T$ is the $(n_{il} \times 2)$-dimensional matrix of response variable within the i-th second level group, within the $l-$th third level group[3],
- $\mathbf{X}_{il}$ is the $(n_{il} \times (P+1))$-dimensional matrix of covariates of fixed effects,
- $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \end{pmatrix}$ is the $((P+1) \times 2)$-dimensional matrix of coefficients of $\mathbf{X}$,
- $\mathbf{Z}_{il}$ is the $(n_{il} \times (R+1))$-dimensional matrix of covariates of random effects,

[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.

[2] In subscript of each variable/parameter, we indicate by the number before the coma whether the variable/parameter is referred to the first or the second response variable (for example, $y_{1,jil}$ and $y_{2,jil}$ are the j-th first and second response variables within (level 2)-group i, that is within (level 3)-group l, 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.

- $\mathbf{c}_{il} = \begin{pmatrix} \mathbf{c}_{1,il} & \mathbf{c}_{2,il} \end{pmatrix}$ is the $((R+1) \times 2)$-dimensional matrix of coefficients of $\mathbf{Z}_{il}$,
- $\boldsymbol{\alpha}_l = \begin{pmatrix} \alpha_{1,l} & \alpha_{2,l} \end{pmatrix}$ is the $(L \times 2)$-dimensional matrix of intercepts of the random effects related to the highest level of the hierarchy,
- $\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_{1,il} & \boldsymbol{\epsilon}_{2,il} \end{pmatrix}$ is the $(n_{il} \times 2)$-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 would consider a three-levels hierarchy: students as level 1, classes as level 2 and schools as level 3. In so doing, thanks to the random intercept $\boldsymbol{\alpha}_l$ at school level, the random 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. Nonetheless, since the available INVALSI data, that will be presented in Section 2, regard classes that are all nested within different schools (each class corresponds to a different school), we can not consider three different levels in our application, but we consider students nested only within classes. Therefore, the three-level model can be reduced to the following two-level model:

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

In the parametric framework of bivariate linear mixed-effects models, the coefficients of random effects 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 & Bates, 2000). This parametric distribution implies that, for each group i, the model estimates the coefficients $\mathbf{c}_i = (c_{i1}, \dots, c_{i(R+1)})$ for the $(R+1)$ covariates of the random effects, meaning that the covariates of random effects are allowed to have N different associations to the response variables across the N 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 $\mathbf{c}_i$ to follow a bivariate discrete distribution P^*, assuming $M \times K$ mass points $(\mathbf{c}_{11}, \dots, \mathbf{c}_{MK})$, 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 $m = 1, \dots, M$ and $k = 1, \dots, K$, 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 modeling 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 mixed-effects model takes the following form:

$$\begin{gathered}\mathbf{Y}_i = \begin{pmatrix}\mathbf{y}_{1,i} & \mathbf{y}_{2,i}\end{pmatrix} = \mathbf{X}_i \begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix}^T + \mathbf{Z}_i \begin{pmatrix}\mathbf{c}_{1,m} \\ \mathbf{c}_{2,k}\end{pmatrix}^T + \boldsymbol{\epsilon}_i \\ i = 1, \ldots, N \qquad m = 1, \ldots, M \qquad k = 1, \ldots, K \\ \boldsymbol{\epsilon}_i^T = \begin{pmatrix}\boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i}\end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind.\end{gathered} \tag{3}$$

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 (3) reduces to:

$$\begin{gathered}\mathbf{Y}_i = \begin{pmatrix}\mathbf{y}_{1,i} & \mathbf{y}_{2,i}\end{pmatrix} = \begin{pmatrix}c_{1,1m} \\ c_{2,1k}\end{pmatrix}^T \mathbf{1} + \sum_{p=1}^{P} \mathbf{x}_{ip} \begin{pmatrix}\boldsymbol{\beta}_{1p} \\ \boldsymbol{\beta}_{2p}\end{pmatrix}^T + \mathbf{z}_i \begin{pmatrix}c_{1,2m} \\ c_{2,2k}\end{pmatrix}^T + \boldsymbol{\epsilon}_i \\ i = 1, \ldots, N \qquad m = 1, \ldots, M \qquad k = 1, \ldots, K \\ \boldsymbol{\epsilon}_i^T = \begin{pmatrix}\boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i}\end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind.\end{gathered} \tag{4}$$

where $\mathbf{1}$ 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 P^* that belongs to the class of all probability measures on $\mathcal{R}^4$. P^* can then be interpreted as the mixing distribution that generates the density of the stochastic model in (4). The ML estimator $\hat{P}^*$ of P^* can be obtained following the theory of mixture likelihoods in (Lindsay et al., 1983a, 1983b), as explained in (Masci et al., 2019). The ML estimator of the random effects distribution can be expressed as a set of points $(\mathbf{c}_{11}, \ldots, \mathbf{c}_{MK})$ and a set of wights $(w_{11}, \ldots, w_{MK})$, 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 $\boldsymbol{\beta}$, $(\mathbf{c}_{11}, \ldots, \mathbf{c}_{MK})$, $(w_{11}, \ldots, w_{MK})$ and $\boldsymbol{\Sigma}$ is performed through the maximization of the likelihood function, mixture by the discrete distribution of random effects,

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

$$L(\boldsymbol{\beta}, \mathbf{c}_{mk}, \boldsymbol{\Sigma}|\mathbf{y}) = \sum_{m=1}^{M}\sum_{k=1}^{K} \frac{w_{mk}}{\sqrt{|det(2\pi\boldsymbol{\Sigma})|^J}} \times$$
$$\times \exp\left\{ \sum_{i=1}^{N}\sum_{j=1}^{n_i} -\frac{1}{2} \begin{pmatrix} y_{1,ij} - c_{1,1m} - \sum_{p=1}^{P} \beta_{1p} x_{1p,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \sum_{p=1}^{P} \beta_{2p} x_{2p,ij} - c_{2,2k} z_{2,ij} \end{pmatrix}^T \boldsymbol{\Sigma}^{-1} \right. \tag{5}$$
$$\left. \begin{pmatrix} y_{1,ij} - c_{1,1m} - \sum_{p=1}^{P} \beta_{1p} x_{1p,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \sum_{p=1}^{P} \beta_{2p} x_{2p,ij} - c_{2,2k} z_{2,ij} \end{pmatrix} \right\}$$

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

3 Simulation study

In this section, we test the performance of the BSPEM algorithm simulating four situations in which the two response variables are related to each other in four different ways, facing both structural correlation/uncorrelation between the subpopulations distributions and correlation/uncorrelation between the errors of the linear model.
We generate $1,000$ bivariate observations that are nested within 100 groups in the following way:

$$\begin{pmatrix} \mathbf{y}_{1,i} & \mathbf{y}_{2,i} \end{pmatrix} = \begin{pmatrix} c_{1,1m} \\ c_{2,1k} \end{pmatrix}^T + \mathbf{x}_i \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}^T + \mathbf{z}_i \begin{pmatrix} c_{1,2m} \\ c_{2,2k} \end{pmatrix}^T + \boldsymbol{\epsilon}_i$$
$$i = 1, \ldots, 100 \qquad m = 1, \ldots, M \qquad k = 1, \ldots, K \tag{6}$$
$$\boldsymbol{\epsilon}_i^T = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind.$$

in which we set $M = 3$ and $K = 2$. Without loss of generality, 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 $\mathbf{x}$, $\mathbf{z}$ and $\boldsymbol{\epsilon}$ in the following way[6]:

$$\begin{aligned} \mathbf{z}_i &\sim \mathcal{N}(0.10, 0.4^2) \quad i = 1, \ldots, 33 \\ \mathbf{z}_i &\sim \mathcal{N}(0.12, 0.4^2) \quad i = 34, \ldots, 66 \\ \mathbf{z}_i &\sim \mathcal{N}(0.08, 0.4^2) \quad i = 67, \ldots, 100 \end{aligned} \tag{7}$$

[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]Again, different choices of values for variables $\mathbf{x}$ and $\mathbf{z}$ are possible and they are also allowed to be different between first and second response variables (i.e. $\mathbf{x}_{1,i} \neq \mathbf{x}_{2,i}$).

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

Table 1: Set of parameters used in Eq. (6) to simulate data. The intercepts and the coefficients of $\mathbf{z}$ differ across subpopulations, while the coefficients β of x are fixed. Colors 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).

$$\begin{aligned} \mathbf{x}_i &\sim \mathcal{N}(0.30, 0.4^2) \quad i = 1, \dots, 33 \\ \mathbf{x}_i &\sim \mathcal{N}(0.28, 0.4^2) \quad i = 34, \dots, 66 \\ \mathbf{x}_i &\sim \mathcal{N}(0.27, 0.4^2) \quad i = 67, \dots, 100 \end{aligned} \tag{8}$$

and

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}_2\left(\mathbf{0}, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad i = 1, \dots, 100. \tag{9}$$

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

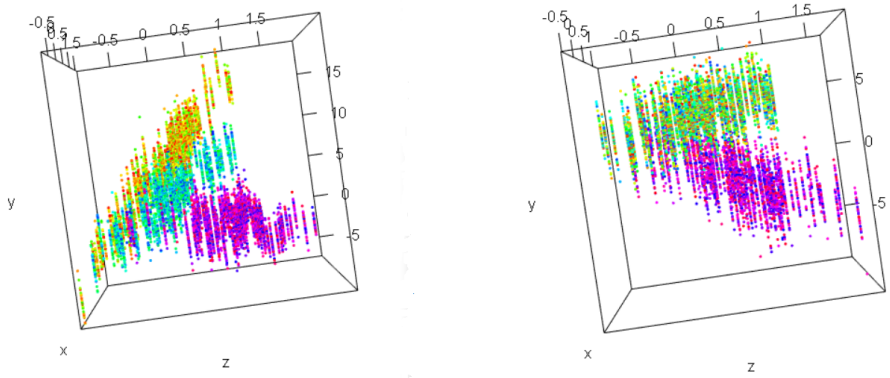

Figure 1: Data simulated with the set of parameters reported in Table 1 and values of $\mathbf{x}$, $\mathbf{z}$ and $\boldsymbol{\epsilon}$ defined in Eq. (7), (8) and (9) respectively. Figure on the left panel represents the first response and figure on the right panel represents the second one. It 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.

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. (9) 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 `tollR` = `tollF` = 10^{-2} (see Algorithm 1 in Appendix). We repeat the simulation for 100 runs. On average, the algorithm converges in 6 iterations and it always identifies 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 W and of $\boldsymbol{\Sigma}$, averaged over the 100 runs, are the following:

	First response parameters	Second response parameters
$i = 1, \ldots, 33$	$\hat{c}_{1,11} = 4.99723$	$\hat{c}_{2,11} = 3.01097$
	$(MSE_{1,11} = 0.00043)$	$(MSE_{2,11} = 0.00024)$
	$\hat{c}_{1,21} = 10.00621$	$\hat{c}_{2,21} = 1.00384$
	$(MSE_{1,21} = 0.00249)$	$(MSE_{2,21} = 0.00091)$
	$\hat{\beta}_1 = 2.99822$	$\hat{\beta}_2 = 1.99856$
	$(MSE_{\beta_1} = 0.00059)$	$(MSE_{\beta_2} = 0.00065)$
$i = 34, \ldots, 66$	$\hat{c}_{1,12} = 2.01128$	$\hat{c}_{2,11} = 3.01066$
	$(MSE_{1,12} = 0.00037)$	$(MSE_{2,11} = 0.00024)$
	$\hat{c}_{1,22} = 4.92278$	$\hat{c}_{2,21} = 1.01334$
	$(MSE_{1,22} = 0.00187)$	$(MSE_{2,21} = 0.00091)$
	$\hat{\beta}_1 = 2.99923$	$\hat{\beta}_2 = 1.99459$
	$(MSE_{\beta_1} = 0.00059)$	$(MSE_{\beta_2} = 0.00065)$
$i = 67, \ldots, 100$	$\hat{c}_{1,13} = 0.00645$	$\hat{c}_{2,12} = -0.00768$
	$(MSE_{1,13} = 0.00195)$	$(MSE_{2,12} = 0.00065)$
	$\hat{c}_{1,23} = -2.00531$	$\hat{c}_{2,22} = -2.99967$
	$(MSE_{1,23} = 0.00203)$	$(MSE_{2,22} = 0.00182)$
	$\hat{\beta}_1 = 2.99948$	$\hat{\beta}_2 = 1.99493$
	$(MSE_{\beta_1} = 0.00059)$	$(MSE_{\beta_2} = 0.00065)$

Table 2: Values of the parameters of Eq. (6) 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). 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).

$$\hat{W} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix} \quad \hat{\Sigma} = \begin{pmatrix} 1.0012 & 0.0001 \\ 0.0001 & 0.9996 \end{pmatrix} \tag{10}$$

$$MSE_{\Sigma} = \begin{pmatrix} 0.0002 & 0.0001 \\ 0.0001 & 0.0003 \end{pmatrix}. \tag{11}$$

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 $\hat{\mathbf{\Sigma}}$, the covariance is correctly estimated as null and the two estimated variances are also close to 1.

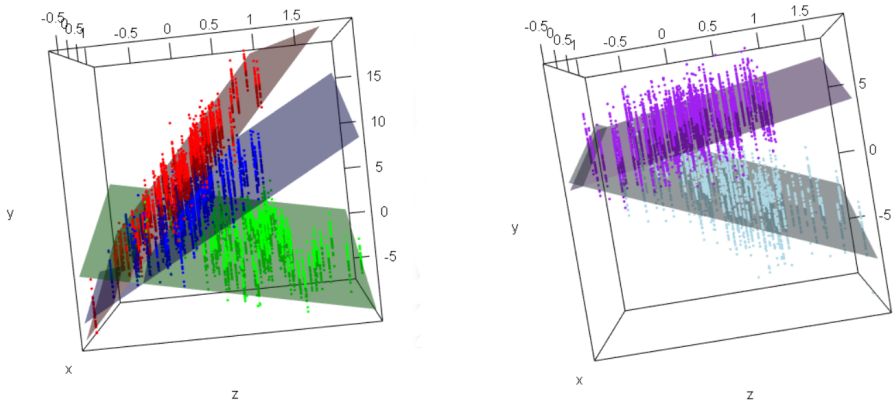

Figure 2: Simulated data with the regression planes identified by the BSPEM algorithm in one of the 100 runs. Colors represent the different subpopulations: three for the first response (figure on the left panel) and two for the second response (figure on the right panel). The estimated parameters of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

The case just shown represents only 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 four different situations:

- Case 1: structural correlation among subpopulations of the two response variables and independence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ (case seen above);
- Case 2: structural correlation among subpopulations of the two response variables and dependence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$;
- Case 3 : not structural correlation among subpopulations of the two response variables and independence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$;
- Case 4: not structural correlation among subpopulations of the two response variables and dependence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$.

In order to avoid a structural correlation among the subpopulations of the two response variables (Case 3 and 4), 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 c_{mk} between the two responses. In order to impose the dependence among the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ (Case 2 and 4), we set the covariance of the variance/covariance matrix $\boldsymbol{\Sigma}$ equal to 0.5. In particular, we set $\Sigma = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$.

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 $D = 1$, $\texttt{tollR} = \texttt{tollF} = 10^{-2}$ (see Algorithm 1 in Appendix). The algorithm is able to identify the correct subpopulations distribution in all the four situations. The visualization of the results in all the four cases is similar to the one shown in Figure 2 and the estimates of the parameters $\mathbf{c}_{m,k}$, for $m = 1, \ldots, 3$ and $k = 1, 2$ and $\boldsymbol{\beta}$ in the four cases are in line with the ones shown in Table 2. What changes across the four cases are the estimates of the weights matrices W and of $\boldsymbol{\Sigma}$, whose means over the 100 runs are shown in Table 3.

	Structural correlation among subpopulations	Not structural correlation among subpopulations
$\boldsymbol{\epsilon}_1 \not\perp\!\!\!\perp \boldsymbol{\epsilon}_2$	$\hat{W} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix}$ $\hat{\Sigma} = \begin{pmatrix} 0.5006 & 0.5006 \\ 0.5006 & 0.5007 \end{pmatrix}$ $MSE_\Sigma = \begin{pmatrix} 0.0003 & 0.0001 \\ 0.0001 & 0.0004 \end{pmatrix}$	$\hat{W} = \begin{pmatrix} 0.25 & 0.08 \\ 0.21 & 0.12 \\ 0.20 & 0.14 \end{pmatrix}$ $\hat{\Sigma} = \begin{pmatrix} 0.4999 & 0.4998 \\ 0.4998 & 0.4999 \end{pmatrix}$ $MSE_\Sigma = \begin{pmatrix} 0.0003 & 0.0002 \\ 0.0002 & 0.0001 \end{pmatrix}$
$\boldsymbol{\epsilon}_1 \perp\!\!\!\perp \boldsymbol{\epsilon}_2$	$\hat{W} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix}$ $\hat{\Sigma} = \begin{pmatrix} 1.016 & 0.001 \\ 0.001 & 0.969 \end{pmatrix}$ $MSE_\Sigma = \begin{pmatrix} 0.0002 & 0.0001 \\ 0.0001 & 0.0004 \end{pmatrix}$	$\hat{W} = \begin{pmatrix} 0.23 & 0.10 \\ 0.22 & 0.12 \\ 0.21 & 0.12 \end{pmatrix}$ $\hat{\Sigma} = \begin{pmatrix} 0.993 & 0.009 \\ 0.009 & 1.023 \end{pmatrix}$ $MSE_\Sigma = \begin{pmatrix} 0.0005 & 0.0001 \\ 0.0001 & 0.0003 \end{pmatrix}$

Table 3: Estimates of the weights matrix W and of the variance/covariance matrix $\boldsymbol{\Sigma}$ (with its MSE computed over the 100 runs) of model in Eq. (6) for the four different cases of values of $\mathbf{c}_{mk}$ and $\boldsymbol{\epsilon}$.

From Table 3, we see that the model is completely 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 W) from the correlation imposed by dependent errors (showed in $\boldsymbol{\Sigma}$). In cases 3 and 4 (last column in Table 3), where we do not impose a structural correlation among subpopulations, the distribution of the weights , less than small variations, is uniformly distributed on the mass points.

The only parameter that significantly influences the results of the BSPEM algorithm is the threshold distance D (see Algorithm 1 in Appendix). 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].

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

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 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 $1,082$ 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). 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[9]. 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 Section 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, personal information of the teacher (age, education, gender), teaching practices and available materials in the class. Table 5 reports teacher and class level variables that we consider, following suggestions derived from the literature about school effectiveness (David, Teddlie, & Reynolds, 2000), with their explanation.

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

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	–	–	–

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

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

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

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

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

The variables shown in Table 5 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 & Resh, 1986, 2018; Belfi, Goos, De Fraine, & Van Damme, 2012; Wayne & 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 & Goldhaber, 1997; Schwerdt & Wuppermann, 2011; Bietenbeck, 2014; De Witte & 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.

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:

$$\begin{gathered}\mathbf{Y}_i = \begin{pmatrix} c_{1,1m} \\ c_{2,1k} \end{pmatrix}^T \mathbf{1} + \sum_{p=1}^{P} \mathbf{x}_{ip} \begin{pmatrix} \boldsymbol{\beta}_{1p} \\ \boldsymbol{\beta}_{2p} \end{pmatrix}^T + \mathbf{z}_i \begin{pmatrix} c_{1,2m} \\ c_{2,2k} \end{pmatrix}^T + \boldsymbol{\epsilon}_i \\ i = 1, \ldots, N \quad m = 1, \ldots, M, \quad k = 1, \ldots, K \\ \boldsymbol{\epsilon}_i^T = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind.\end{gathered} \tag{12}$$

where i is the class index and N is the total number of classes. $\mathbf{Y}_i = \begin{pmatrix} \texttt{math8}_i & \texttt{read8}_i \end{pmatrix}$ 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: $D_1 = D_2 = 0.3$, $\tilde{w}_1 = \tilde{w}_2 = 0.01$, `tollR` = `tollF` = 10^{-2}, `it`=40, `itmax`=20, `it1`=20 (see Algorithm 1 in Appendix). The algorithm converges in 30 iterations and identifies $M = 5$ mass points for the random effects 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 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 6.

First response variable						
	$\hat{c}_{1,1}$ (intercept)	$\hat{c}_{1,2}$ (`math5`)	$\hat{w}_1$ (weight)	$\hat{\beta}_{11}$ (`ESCS`)	$\hat{\beta}_{12}$ (`gender`)	$\hat{\beta}_{13}$ (`immigrant`)
m=1	0.295	0.719	0.458			
m=2	−0.181	0.464	0.384			
m=3	0.762	0.463	0.025	0.089	−0.055	0.048
m=4	−1.301	0.112	0.064			
m=5	0.366	0.291	0.069			
Second response variable						
	$\hat{c}_{2,1}$ (intercept)	$\hat{c}_{2,2}$ (`read5`)	$\hat{w}_2$ (weight)	$\hat{\beta}_{21}$ (`ESCS`)	$\hat{\beta}_{22}$ (`gender`)	$\hat{\beta}_{23}$ (`immigrant`)
k=1	−2.848	−0.101	0.019			
k=2	−0.622	0.262	0.095	0.095	0.219	−0.083
k=3	−1.556	0.188	0.018			
k=4	0.054	0.544	0.868			

Table 6: Estimates of the parameters of Eq. (12) obtained by the BSPEM algorithm, related to the two response variables. The coefficients $\boldsymbol{\beta}$ of the fixed effects do not change across subpopulations.

$\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are the coefficients of fixed effects 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 coefficients of random effects 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 higher 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), Figure 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 6 and the regression lines in Figure 3, it is possible to

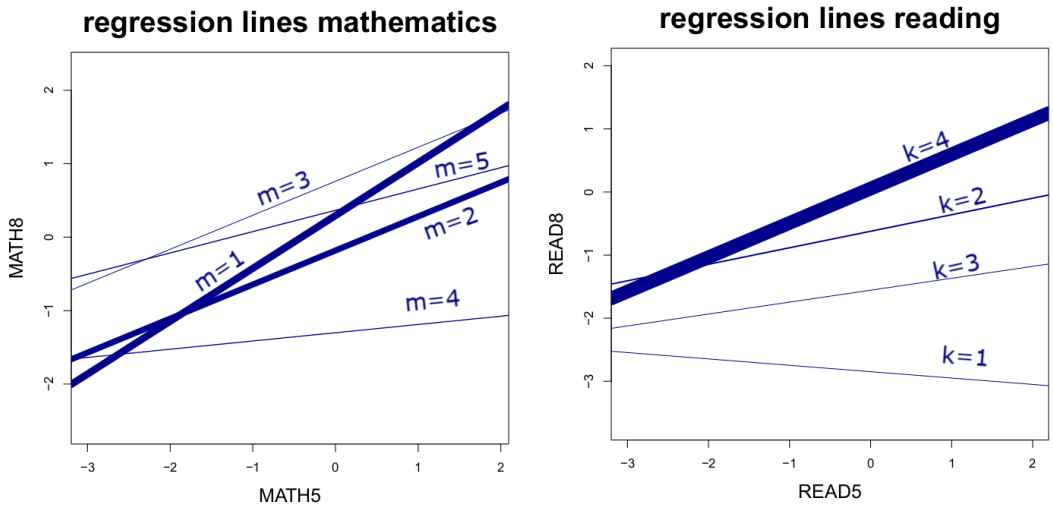

Figure 3: Regression planes projected on the 2-dimensional plane identified by the answer variable and the random covariate, identified by the parameters of Eq. (12) estimated by the BSPEM algorithm and whose parameters are shown in Table 6. Panel on the left reports the results for the first response, 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.

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 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[10], reported in Table 7. 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 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 W and the variance/covariance matrix $\boldsymbol{\Sigma}$ are estimated as follows:

$$\hat{W} = \begin{pmatrix} 0.0000 & 0.0007 & 0.0003 & 0.4571 \\ 0.0054 & 0.0518 & 0.0047 & 0.3220 \\ 0.0022 & 0.0000 & 0.0023 & 0.0204 \\ 0.0068 & 0.0312 & 0.0082 & 0.0179 \\ 0.0043 & 0.0111 & 0.0029 & 0.0507 \end{pmatrix} \quad \hat{\Sigma} = \begin{pmatrix} 0.455 & 0.183 \\ 0.183 & 0.451 \end{pmatrix}. \tag{13}$$

The covariance and the correlation among the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are 0.183 and 0.404, respectively.

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

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

Table 7: Average values of some student level variables used in the analysis, across the identified subpopulations (five for mathematics and four for reading).

Considering the two marginal distributions of the class effects, we observe from Table 6 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 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.
- 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 8 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

	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	$S_3(+-)$		$S_3(+-)$	

Table 8: 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.

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 5. 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, Hastie, & Tibshirani, 2010) and are shown in Table 9.

Variable name	$\boldsymbol{S_{ref}}$	$\boldsymbol{S_2}$	$\boldsymbol{S_3}$	$\boldsymbol{S_4}$
Teachers general questions				
`contacts among maths teachers`	-0.186			
`contacts among reading teachers`	-0.169			
Teachers personal information				
`Maths teacher age`	-0.024			
`Reading teacher age`	-0.003			
`Reading teacher gender (male=1)`				0.249
Only for mathematics teachers				
`main teaching method 'a'`	-0.132			
`teacher written exams`		-1.729		
Only for reading teachers				
`num years of teaching here`	-0.015			
`num reading hours`	-0.016			
`summarize text`	-0.071			
`read newspapaper`		0.610		
Class information and body composition				
% 2^{nd}`-gen immig`	4.519			
`area geo South`	-1.417	0.048		

Table 9: 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.

According to the results of the multinomial logit model shown in Table 9, 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, the geographical area and the percentage of immigrant (second generation) students. Classes where teachers of reading 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. Classes with a high percentage of second generation immigrant students are more likely to belong to S_{ref}, suggesting the positive effect of diversity of class composition (this result is also partially explained by the fact that the percentage of immigrant students in Southern Italy is very low with respect to the one 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.

5 Conclusions

In this paper, we develop a bivariate semi-parametric mixed-effects model, 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/schools. 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 mixed-effects 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/schools), 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/school). 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 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. Σ_{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/school 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/school). 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. 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. The proposed model is already presented in its complete form in this paper, for allowing empirical analyses in this direction.

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.

References

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

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. *Educational research review*, *7*(1), 62–74.

Bietenbeck, J. (2014). Teaching practices and cognitive skills. *Labour Economics*, *30*, 143–153.

Brewer, D. J., & Goldhaber, D. (1997). Why don't schools and teachers seem to matter?: Assessing the impact of unobservables on educational productivity. *Journal of Human Resources, Summer*.

Dar, Y., & Resh, N. (1986). Classroom intellectual composition and academic achievement. *American Educational Research Journal*, *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. *Education Economics*, *22*(1), 3–23.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1. Retrieved from `www.jstatsoft.org/v33/i01/`

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, 505–523.

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

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of political economy*, *100*(1), 84–117.

Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in england 1992–2016:'contextual value-added','expected progress' and 'progress 8'. *British Educational Research Journal*, *43*(2), 193–212.

Lin, L. I., et al. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine*, *19*(2), 255–270.

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

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

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

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35–62.

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

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

Masci, C., Paganoni, A. M., & Ieva, F. (2019). Semiparametric mixed effects models for unsupervised classification of italian schools. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. Retrieved from `https://doi.org/10.1111/rssa.12449`

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, *29*(1), 67–101.

McCulloch, C., Lin, H., Slate, E., & Turnbull, B. (2002). Discovering subpopulation structure with latent class mixed models. *Statistics in medicine*, *21*(3), 417–429.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*(3), 283–301.

Muthén, B. (2004). Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences*, *345*, 368.

Muthén, B., & Asparouhov, T. (2015). Growth mixture modeling with non-normal distributions. *Statistics in Medicine*, *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, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods*, *4*(2), 139.

Palardy, G. J. (2008). Differential school effects among low, middle, and high social class composition schools: A multiple group, multilevel latent growth curve analysis. *School Effectiveness and School Improvement*, *19*(1), 21–49.

Parsons, E., Koedel, C., & Tan, L. (2018). Accounting for student disadvantage in value-added models. *Journal of Educational and Behavioral Statistics*, *4*, 144–179.

Perry, T. (2016). English value-added measures: Examining the limitations of school performance measurement. *British Educational Research Journal*, *42*(6), 1056–1080.

Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, 3–56.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of educational and behavioral statistics*, *20*(4), 307–335.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

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

Schagen, I., & Schagen, S. (2005). Combining multilevel analysis with national value-added data sets - a case study to explore the effects of school diversity. *British Educational Research Journal*, *31*(3), 309–328.

Schwerdt, G., & Wuppermann, A. C. (2011). Is traditional teaching really all that bad? a within-student between-subject approach. *Economics of Education Review*, *30*(2), 365–379.

Strand, S. (1997). Pupil progress during key stage 1: a value added analysis of school effects. *British educational research journal*, *23*(4), 471–487.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Timmermans, A. C., Bosker, R. J., de Wolf, I. F., Doolaard, S., & van der Werf, M. P. (2014). Value added based on educational positions in dutch secondary education. *British Educational Research Journal*, *40*(6), 1057–1082.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis*, *11*, 89–106.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational research*, *73*(1), 89–122.

Wenglinsky, H. (2002). The link between teacher classroom practices and student academic performance. *Education policy analysis archives*, *10*, 12.

Winkler, D. R. (1975). Educational achievement and school peer group composition. *Journal of Human Resources*, 189–204.

Appendix

The EM algorithm for bivariate semi-parametric mixed-effects linear models

The EM algorithm that we propose to estimate the parameters of the model in (4) 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 effects, given the observations and the parameters computed in the previous iteration; and the maximization step (M step) in which we maximize the conditional expectation of the likelihood function. At each iteration, the EM algorithm updates the parameters in order to increase the likelihood in Eq. (5) and it continues until the convergence. The update of the parameters is the following:

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

and

$$\begin{aligned}(\boldsymbol{\beta}^{(up)}, \mathbf{c}_{mk}^{(up)}, \boldsymbol{\Sigma}^{(up)}) = \\ \arg\max_{\beta, \mathbf{c}_{mk}, \boldsymbol{\Sigma}} \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{i=1}^{N} W_{imk}\ \ln p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{c}_{mk})\end{aligned} \tag{17}$$

where

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

and

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

The coefficient W_{imk} represents the probability of $\mathbf{c}_i$ being equal to $\mathbf{c}_{mk}$ conditionally to observations $\mathbf{y}_i$ and given the fixed coefficient $\boldsymbol{\beta}$ and the variance/covariance matrix $\boldsymbol{\Sigma}$. Indeed, since $w_{mk} = p(\mathbf{c}_i = \mathbf{c}_{mk})$, then

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

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

$$\hat{\mathbf{c}}_i = \mathbf{c}_{\tilde{mk}} \quad \text{where} \quad \tilde{mk} = \arg\max_{m,k} W_{imk} \quad i = 1, \ldots, N. \tag{21}$$

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 $\boldsymbol{\beta}$ 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

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

Since we are considering the linear case, the maximization step is done in closed-form[11]. In the second step, we fix the support points of the random effects distribution computed in the previous step and we compute the *arg-max* in Eq. (17) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Again, this step is done in closed-form.

The initialization of the support points of the discrete distribution P^* 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 coefficients of random effects and a starting estimate for the coefficient of the fixed effects, for both the response variables. These parameters are chosen in the following way:

- random effects: 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;

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

- fixed effects: the starting values of $\boldsymbol{\beta}$ 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 effects 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 effects, 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 $\mathbf{c}_{1,(hg)} = \frac{\mathbf{c}_{1,h}+\mathbf{c}_{1,g}}{2}$. 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 $k = 1, \ldots, K$. 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 the threshold value D is the same for the clusters related to the two response variables, but the procedure might lead to different number of mass points M and K. The latter 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 $m = 1, \ldots, M$ and $w_{2,k} \leq \tilde{w}$, for $k = 1, \ldots, K$ or that are not associated to any subpopulation. D and $\tilde{w}$ are two tuning parameters that tune the estimates of the subpopulations. 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. (4) in an iterative way, updating the coefficients related to both fixed and random effects, until convergence or until

it reaches the maximum number of sub-iterations fixed 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 parameters of fixed and random effects 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).

Algorithm 1: EM algorithm for bivariate semi-parametric mixed-effects models

input : Initial estimates for $(\mathbf{c}_{11}^{(0)}, \ldots, \mathbf{c}_{MK}^{(0)})$ and $(w_{11}^{(0)}, \ldots, w_{MK}^{(0)})$, with $M = N$ and $K = N$;
Initial estimates for $\boldsymbol{\beta}^{(0)}$ 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 $m = 1, \ldots, M$, $k = 1, \ldots, K$, $\boldsymbol{\beta}^{(a)}$ and $\Sigma^{(a)}$.

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 $s, t \; \forall s, t = 1, \ldots, M, s \neq t$);

if *(*$DIST1_{st} < D_1$ & $DIST1_{st} = min(DIST1) \quad (\forall s, t = 1, \ldots, M, s \neq t)$*)* **then**
collapse marginal masses s and t to a unique mass point;

if *(*$DIST2_{st} < D_2$ & $DIST2_{st} = min(DIST2) \quad (\forall s, t = 1, \ldots, K, s \neq t)$*)* **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 $w_{1,m}^{(a)} \leq \tilde{w}_1 \quad (\forall m = 1, \ldots, M)$ **then**
delete marginal mass point m;
reparameterize the weights;

if $w_{2,k}^{(a)} \leq \tilde{w}_2 \quad (\forall k = 1, \ldots, K)$ **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 $m = 1, \ldots, M$ and $k = 1, \ldots, K$, $\boldsymbol{\beta}^{(a-1)}$ and $\Sigma^{(a-1)}$, compute the matrix W according to Eq. (20);

update the weights $w_{11}^{(a)}, \ldots, w_{MK}^{(a)}$ according to Eq. (16);

$\boldsymbol{\beta}^{(a,0)} = \boldsymbol{\beta}^{(a-1)}$;
$\Sigma^{(a,0)} = \Sigma^{(a-1)}$;
$\mathbf{c}_{mk}^{(a,0)} = \mathbf{c}_{mk}^{(a-1)}$;
$w_{mk}^{(a,0)} = w_{mk}^{(a-1)}$;

keeping $\boldsymbol{\beta}^{(a,0)}$ and $\Sigma^{(k,0)}$ fixed, update the $M \times K$ support points $\mathbf{c}_{11}^{(a,1)}, \ldots, \mathbf{c}_{MK}^{(a,1)}$ according to Eq. (17);

keeping $\mathbf{c}_{mk}^{(a,1)}, w_{mk}^{(a,0)}$ for $m = 1, \ldots, M$ and $k = 1, \ldots, K$ fixed, update $\boldsymbol{\beta}^{(a,1)}$ and $\Sigma^{(a,1)}$ according to Eq. (17);

j=1;

while
$(|\boldsymbol{\beta}^{(a,j-1)} - \boldsymbol{\beta}^{(a,j)}| \geq tollF \quad or \quad |\Sigma^{(a,j-1)} - \Sigma^{(a,j)}| \geq tollF \quad or \quad |\mathbf{c}_{mk}^{(a,j-1)} - \mathbf{c}_{mk}^{(a,j)}| \geq tollR) \quad \& \quad j \leq itmax$
do
j=j+1;
keeping $\boldsymbol{\beta}^{(a,j-1)}$ and $\Sigma^{(a,j-1)}$ fixed, update the $M \times K$ support points $\mathbf{c}_{11}^{(a,j)}, \ldots, \mathbf{c}_{MK}^{(a,j)}$ according to Eq. (17);
keeping $\mathbf{c}_{mk}^{(a,j)}, w_{mk}^{(a,j-1)}$ for $m = 1, \ldots, M$ and $k = 1, \ldots, K$ fixed, update $\boldsymbol{\beta}^{(a,j)}$ and $\Sigma^{(a,j)}$ according to Eq. (17);

set $\mathbf{c}_{mk}^{(a)} = \mathbf{c}_{mk}^{(a,j)}$ for $m = 1, \ldots, M$ and $k = 1, \ldots, K$, $\boldsymbol{\beta}^{(a)} = \boldsymbol{\beta}^{(a,j)}$, $\Sigma^{(a)} = \Sigma^{(a,j)}$;

estimate subpopulation mk for each group i according to Eq. (21);

if $(\boldsymbol{\beta}^{(a)} - \boldsymbol{\beta}^{(a-1)} < tollF)$ & $(\Sigma^{(k)} - \Sigma^{(k-1)} < tollF)$ & $(\mathbf{c}_{mk}^{(a)} - \mathbf{c}_{mk}^{(a-1)} < tollR)$ **then**
$conv1 = 1$;

a= a+1;

MOX Technical Reports, last issues

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

20/2019 Martino, A.; Guatteri, G.; Paganoni, A.M.
Hidden Markov Models for multivariate functional data

21/2019 Martino, A.; Guatteri, G.; Paganoni, A.M.
Hidden Markov Models for multivariate functional data

22/2019 Gigante, G.; Sambataro, G.; Vergara, C.
Optimized Schwarz methods for spherical interfaces with application to fluid-structure interaction

23/2019 Laurino, F; Zunino, P.
Derivation and analysis of coupled PDEs on manifolds with high dimensionality gap arising from topological model reduction

18/2019 Delpopolo Carciopolo, L.; Cusini, M.; Formaggia, L.; Hajibeygi, H.
Algebraic dynamic multilevel method with local time-stepping (ADM-LTS) for sequentially coupled porous media flow simulation

19/2019 Torti, A.; Pini, A.; Vantini, S.
Modelling time-varying mobility flows using function-on-function regression: analysis of a bike sharing system in the city of Milan.

17/2019 Antonietti,P.F.; De Ponti, J.; Formaggia, L.; Scotti, A.
Preconditioning techniques for the numerical solution of flow in fractured porous media

14/2019 Antonietti, P.F.; Facciolà, C; Verani, M.
Mixed-primal Discontinuous Galerkin approximation of flows in fractured porous media on polygonal and polyhedral grids

15/2019 Brandes Costa Barbosa, Y. A.; Perotto, S.
Hierarchically reduced models for the Stokes problem in patient-specific artery segments

16/2019 Antonietti, P.F.; Houston, P.; Pennesi, G.; Suli, E.
An agglomeration-based massively parallel non-overlapping additive Schwarz preconditioner for high-order discontinuous Galerkin methods on polytopic grids