Clustering students according to their proficiency: a comparison between different approaches based on item response theory models

Rosa Fabbricatore, Francesco Palumbo

1. Introduction

Evaluating learners' competencies is a crucial concern in education, and home and classroom structured tests represent an effective assessment tool. Structured tests consist of sets of items that can refer to several abilities or more than one topic. Several statistical approaches allow evaluating students considering the items in a multidimensional way, accounting for their structure. According to the evaluation's ending aim, the assessment process assigns a final grade to each student or clusters students in homogeneous groups according to their level of mastery and ability. The latter represents a helpful tool for developing tailored recommendations and remediatiodddns for each group (Davino et al., 2020; Fabbricatore et al., 2021). At this aim, latent class models represent a reference.

In the item response theory (IRT) paradigm, the multidimensional latent class IRT models, releasing both the traditional constraints of unidimensionality and continuous nature of the latent trait, allow detecting sub-populations of homogeneous students according to their proficiency level also accounting for the multidimensional nature of their ability (Bartolucci et al., 2014). Moreover, the semi-parametric formulation leads to several advantages in practice: It avoids normality assumptions that may not hold and reduces the computation demanding.

However, when the interest is to accurately estimate the individual level of ability in addition to the clustering purpose, a two-step approach could be used.

In this vein, this study compares the results of the multidimensional latent class IRT models with those obtained by a two-step procedure, which consists of firstly modeling a set of unidimensional IRT models to estimate students' ability in each knowledge domain and then applying a clustering algorithm to classify students accordingly. Regarding the latter, parametric and non-parametric approaches were considered. In particular, the k-means clustering algorithm (MacQueen, 1967), the Gaussian mixture model-based clustering (McLachlan and Peel, 2000), and the archetypal analysis (Cutler and Breiman, 1994) were implemented.

The aim is to investigate similarities and differences in groups detection and students' classification. Indeed, describing students' profiles according to a set of reference groups can take many forms, depending on the adopted approach and estimation procedure.

2. Data and procedure

Data refer to the $N = 944$ subjects involved in the admission test for the degree course in psychology exploited in 2014 at the University of Naples Federico II.

The following five different domains represent the knowledge dimensions assessed by the admission test: Humanities (30 items), Reading (30 items), Mathematics (10 items), Science (10 items), and English (20 items). Correct answers receive one credit and are coded with 1, whereas blank and wrong answers receive no credit and are coded as 0.

Rosa Fabbricatore, University of Naples Federico II, Italy, rosa.fabbricatore@unina.it, 0000-0002-4056-4375
Francesco Palumbo, University of Naples Federico II, Italy, fpalumbo@unina.it

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Rosa Fabbricatore, Francesco Palumbo, *Clustering students according to their proficiency: a comparison between different approaches based on item response theory models*, pp. 43-48, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-461-8.09, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci (edited by), *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the on-site conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-461-8 (PDF), DOI 10.36253/978-88-5518-461-8

Firstly, we carried out the multidimensional latent class IRT model to cluster subjects into classes as homogeneous as possible according to their abilities, concurrently accounting for the multidimensional structure of the data. Secondly, we implemented three two-step procedures exploiting the k-means algorithm, the Gaussian mixture modeling, and the archetypal analysis, respectively. Finally, we compared the different approaches employing a graphic example and evaluated their agreement through the Adjusted Rand Index (ARI; Hubert and Arabie, 1985). The ARI is a commonly used measure to evaluate distances in clustering. It allows comparing a partition with another one on the same elements or with external criteria. Index computation is based on the number of pairs of elements that are allocated in the same (or different) cluster in both partitions (agreements) and the number of pairs of elements that are placed in the same cluster in one partition but in different clusters in the other (disagreements). The ARI values range from 0 (random partitioning) to 1 (partitions perfect agreement).

3. Statistical method

Methods we compared in this study exploit IRT models for students' ability estimation (see Bartolucci et al. (2019) for a review on the IRT models). In more detail, we considered the two-parameter logistic (2PL) IRT parametrization, where the parameters of guessing and ceiling are constrained to be equal to 0. Thus the probability of correct response depends only on the discrimination and difficulty item parameters and the student's ability. More formally, the probability that the subject s correctly answers the dichotomously-scored item i (with $i = 1, \ldots, I$) can be expressed as follows:

$$P(X_{si} = 1|\theta_s, a_i, b_i) = \frac{e^{a_i(\theta_s - b_i)}}{1 + e^{a_i(\theta_s - b_i)}}, \tag{1}$$

where X_{si} is the response of the subject s at the item i with realization $x_{si} \in [0, 1]$, $\theta_s \in R$ is the ability of the subject s, $a_i \in R$ is the item discrimination parameter, and $b_i \in R$ represents the item difficulty. It is worth noting that traditional IRT models are ground on three main assumptions: unidimensionality, monotonicity, and local independence. Moreover, the latent trait is described by a continuous normal probability distribution.

Within this theoretical paradigm, the multidimensional latent class IRT models represent a semi-parametric formulation of the traditional IRT models, allowing releasing both the constraints of unidimensionality and the continuous nature of the latent trait. This extension is particularly useful for detecting sub-populations of homogeneous students according to their ability level.

Since we defined the ability as a multidimensional latent trait, each subject is described by the ability vector $\mathbf{\Theta}_s = (\Theta_{s1}, \Theta_{s2}, \ldots, \Theta_{sD})'$ where D is the number of considered dimensions. Following the between-item multidimensional formulation, each item measures only one dimension, and thus items are divided into different subsets I_d with $d = 1, 2, \ldots, D$.

Moreover, according to the semi-parametric formulation, each latent trait have a discrete distribution with $\xi_1, \ldots, \xi_k$ support points defining k latent classes with weights $\pi_1, \ldots, \pi_k$. The main assumption is that subjects in the same latent class share common levels of the latent trait. The generic class weight π_c (with $c = 1, \ldots, k$) represents the probability of belonging to class c and can be expressed as $\pi_c = P(\Theta_s = \xi_c)$ with $\sum_{c=1}^{k} \pi_c = 1$ and $\pi_c \geq 0$.

Accordingly, the manifest distribution of the response vector $\mathbf{X} = (X_1, \ldots, X_I)'$ can be formalized as:

$$P(\mathbf{X} = \mathbf{x}) = \sum_{c=1}^{k} P(\mathbf{X} = \mathbf{x}|\mathbf{\Theta} = \xi_c)\pi_c = \sum_{c=1}^{k} \prod_{d=1}^{D} \prod_{i \in I_d} P(X_i = x_i|\Theta_d = \xi_{cd})\pi_c, \tag{2}$$

where $P(X_i = x_i|\Theta_d = \xi_{cd})$ It is herein specified according to the 2PL parameterization.

The number of classes k can be derived from theoretical assumptions or by comparing the model fit measures at different values of k. Each unit was assigned to the class that corresponds to the highest probability of belonging.

The estimation of the model parameters is usually based on the Maximum Marginal Likelihood (MML) approach. In the specific case of the latent class formulation, the Expectation-Maximization (EM) algorithm is used (Dempster et al., 1977). The estimation process is performed through the R packages `mirt` (Chalmers, 2012) and `MultiLCIRT` (Bartolucci et al., 2014) for the parametric and semi-parametric IRT formulation, respectively.

As stated before, the latent class IRT models allow removing parametric assumptions that may not hold and make the estimation process computationally demanding. Moreover, they are more flexible than the parametric formulation when the main aim is clustering individuals. However, this semi-parametric formulation provides a less accurate estimate of the individual level ability than the continuous one.

Regarding the clustering algorithms applied on the ability estimates, a very brief description below. The *k-means* produces a hard clustering changing the data partition at each step taking into account the Euclidean distance of each point from the cluster centers. It is one of the most used algorithms in cluster analysis mainly due to its ease of implementation and interpretation. Nevertheless, the k-means algorithm works well only when dealing with spherical clusters and no outliers are present in the data set. Firstly, accounting only for clusters' centroids is not suitable enough to properly detect subpopulations that also have covariance parameters significantly different. Secondly, centroids could be dragged by outliers.

Overcoming these issues, the *Gaussian mixture model* provides a model-based clustering allowing to detect differences between sub-populations that share the same (Gaussian) distribution but have one or more different vectors of parameters; thus, these models estimate a specific covariance matrix for each cluster and better manage the presence of outliers. On the other hand, they could entail the risk of overparameterization: increasing model complexity does not guarantee a better solution to the classification problem.

Compared to the methods mentioned above, the archetypal analysis allows more separate groups, detecting extreme representative observations that differ from each other as much as possible. Consequently, it approximates each point in a dataset as a convex combination of this set of extreme data points, called archetypes, lying on the convex hull of the data. Conversely, drawbacks reside in its computation costs, especially as the number of observations increases.

The corresponding R packages used to carry out the analyses were `stats`, `mclust` (Scrucca et al., 2016), and `archetypes` (Eugster, 2009).

4. Results

A set of multidimensional latent class IRT models with a different number of latent classes k were estimated. Basing on the Bayesian information criterion (BIC; Schwarz, 1978), we chose the model with $k = 3$ as the best one for describing our data. Looking at support points, we notice that latent classes are decreasing ordered according to the students' proficiency levels in all the considered domains (see Table 1). In particular, Class 1 encompasses students with poor performance in all the six domains; Class 2 includes students with low performance in Humanities, Math, Science and English, and high performance in Reading; Class 3 consists of students with a good performance in all the domains except for Humanities for which they achieved an average performance. Class weights indicate that Class 2 (moderate ability) is the largest one ($\pi_2 = 0.48$), followed by Class 1 (higher ability; $\pi_1 = 0.39$).

This result was compared with students' classifications obtained by a two-step procedure.

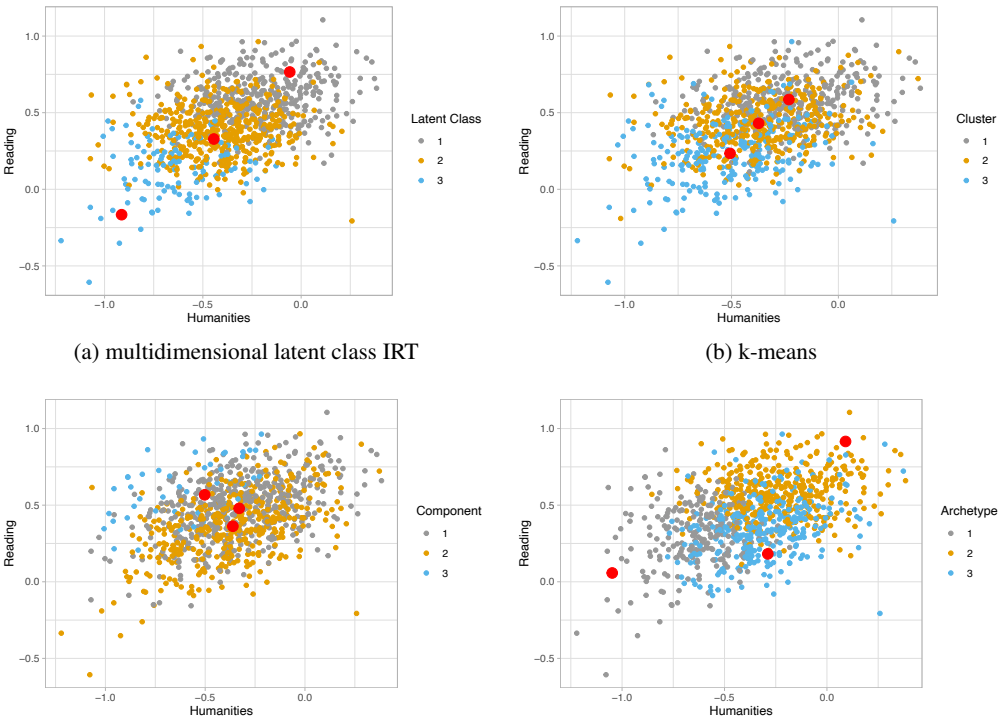
As stated before, the comparison involved different clustering algorithms that were carried out on the students' ability estimates provided by the set of unidimensional IRT models (see Table 1). It is worth noting that the number of classes was imposed equal to $k = 3$ in all the clustering procedure for comparison purposes.

Table 1: Standardized support points (Latent Class IRT), centroids (K-means), component means (Gaussian Mixture Model), archetypes (Archetypal Analysis), and class weights (π_c) for each clustering approaches.

		Latent Class IRT	K-means	Gaussian Mixture Model	Archetypal Analysis
Class 1	Humanities	-0.91	-0.52	-0.50	-0.54
	Reading	-0.17	0.23	0.57	0.01
	Math	-0.81	-0.36	-0.08	-0.87
	Science	-0.75	-0.27	0.11	-0.59
	English	-1.22	-0.68	-0.28	-1.22
Class 2	Humanities	-0.44	-0.38	-0.36	-1.01
	Reading	0.32	0.42	0.36	0.26
	Math	-0.24	-0.15	-0.36	0.22
	Science	-0.10	0.03	-0.09	-0.18
	English	-0.21	-0.05	-0.06	-0.17
Class 3	Humanities	-0.06	-0.24	-0.33	0.21
	Reading	0.76	0.58	0.47	0.92
	Math	0.35	0.15	0.15	0.47
	Science	0.51	0.28	0.18	0.79
	English	0.72	0.53	0.16	1.29
Class weight	π_1	*0.13*	*0.21*	*0.08*	*0.35*
	π_2	*0.48*	*0.42*	*0.43*	*0.24*
	π_3	*0.39*	*0.37*	*0.49*	*0.41*

The example reported in Figure 1, showing only two of the five ability dimensions for simplicity and lack of space, allows depicting differences in students allocation due to the considered clustering approaches. As can be guessed from the picture, the multidimensional latent class IRT model reached the strongest agreement in terms of classification when the k-means algorithm was implemented ($ARI = 0.53$). The confusion matrix showed that the main difference resided in a higher allocation rate in class 2 rather than in class 1 for the multidimensional latent class IRT model compared to the approach based on k-means. A weaker agreement was found with the archetypal analysis ($ARI = 0.39$), whereas the lowest one was reported with the parametric approach based on Gaussian mixture modelling ($ARI = 0.09$).

Notice that in addition to the ability estimates in Table 1, the considered clustering approaches also differ for the allocation procedure that strongly influences the level of agreement between the partitions and, consequently, the ARI.



(a) multidimensional latent class IRT

(b) k-means

(c) Gaussian mixture model

(d) archetypal analysis

Figure 1: Students allocation based on different clustering approaches. X-axes and y-axes refer to students' ability estimation through the multidimensional IRT model in Humanities and Reading, respectively. According to the considered method, red points indicate: (a) standardized support points, (b) centroids, (c) component means, and (d) archetypes.

5. Conclusion

The study provides a useful insight in understanding dissimilarities between different approaches used for clustering purposes. Assuming as a matter of the fact that the adequacy of a method mainly depends on research goals and thus that there is not the best one in absolute terms, we compared different approaches illustrating which of the clustering algorithm we considered in the two-step procedure provides results more similar to those obtained by the multidimensional latent class IRT model.

The proposed comparison also invokes the difference between the parametric and semi-parametric formulation of IRT models in practical applications.

Future research should investigate how the considered approaches work when a different data structure holds. Moreover, it would be interesting also to consider differences deriving from classical test theory rather than the IRT paradigm for the ability estimation.

References

Bartolucci, F., Bacci, S., Gnaldi, M. (2014). MultiLCIRT: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis*, **71**, pp. 971–985.

Bartolucci, F., Bacci, S., Gnaldi, M. (2019). *Statistical analysis of questionnaires: A unified*

approach based on R and Stata. Chapman and Hall/CRC, London, (UK).

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, **48**(6), pp. 1–29.

Cutler, A., Breiman, L. (1994). Archetypal analysis. *Technometrics*, **36**(4), pp. 338–347.

Davino, C., Fabbricatore, R., Pacella, D., Vistocco, D., Palumbo, F. (2020). Aleas: a tutoring system for teaching and assessing statistical knowledge, in *Proceedings of the Second Symposium on Psychology-Based Technologies*, eds. O. Gigliotta and M. Ponticorvo, CEUR Workshop Proceedings, 2730.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), pp. 1–22.

Eugster M.J.A., Leisch, F. (2009). From Spider-Man to Hero – Archetypal Analysis in R. *Journal of Statistical Software*, **30**(8), pp. 1–23.

Fabbricatore, R., Parola, A., Pepicelli, G., Palumbo, F. (2021). A latent class approach for advising in learning statistics: implementation in the ALEAS system, in *Proceedings of the First Workshop on Technology Enhanced Learning Environments for Blended Education - The Italian e-Learning Conference 2021*, eds. P. Limone and R. Di Fuccio, CEUR Workshop Proceedings, 2817.

Hubert, L., Arabie, P. (1985). Comparing partitions. *Journal of classification*, **2**(1), pp. 193–218.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, eds. L. Lecam and J. Neyman, Cambridge University Press, Oakland, (CA), pp. 281–297.

McLachlan, G.J., Peel, D. (2000). *Finite mixture models*. Wiley-Interscience, New York, (NY).

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics*, **6**(2), pp. 461–464.

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, **8**(1), pp. 289–317.