# Combining parametric and non-parametric methods for the analysis of complex small big data structures

Iodice D'Enza A., Iannario M., Romano R.

**Key words:** distance learning; joint data reduction; latent variables

## 1 Introduction

The definition of Big Data varies with the domain of application. General characteristics of Big Data are high volume, data-collection rate, variety of both data structures and storage platforms; the impact of one or more of these characteristics depends on the application. Often Big Data are *second-hand*, that is, the available data is not collected for analysis purposes in the first place. Compared to Big Data, survey data have opposite characteristics: in fact, survey data usually come in low volume, and they are thoughtfully collected for analysis purposes [3]. Therefore, survey data are referred to as an example of Small Data [8]: in some applications it is better-off having a low volume of high quality data to study a phenomenon. The definition of Small Big Data [4] refers to data structures that merge (some of) the characteristics of Big Data (high volume, different sources), and Small Data (high quality, carefully collected). Large scale, complex-structured survey data, that can be referred to as Small Big Data, are analysed in the private and public sectors: such data structures are characterised by a large number of polytomous or mixed attributes, by the possible presence of latent groups of observations and of block-wise structure of attributes. We confine our analysis in this framework studying Small Big Data related to Distance Learning (DL), a recent phenomenon that, because of

Iodice D'Enza A., Iannario M.
Dipartimento di Scienze Politiche, Università degli Studi di Napoli Federico II
e-mail: iodicede@unina.it; maria.iannario@unina.it

Romano R.
Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Napoli Federico II
e-mail: rosaroma@unina.it

the Covid-19 pandemic, had a dramatic increase. Measuring the impact of DL on students is of crucial importance; the technical setbacks that jeopardise the learning experience, are relatively easy to identify and to quantify. A more sensitive issue is to study the effects of the DL on students from a social and psychological perspective. In order to investigate the faceted DL impact on students, we considered three different scales proposed and validated in the literature. In particular, we examined the scale proposed by [1] to study the perspective of DL high-educated students; two further scales were considered, the 'student stress scale', proposed and validated by [16], and the 'fear of Covid-19' scale, proposed by [9], that investigates the future career anxiety. The three scales and some respondents' characteristics are collected in a survey concerning high-educated students that consists of four item-blocks. The survey, carried out in 2020, refers to 1592 students from 60 Italian Universities, with University of Naples and University of Bologna being the most represented. The response option for the majority of items is a 4 levels Likert-type scale, ranging from *strongly disagree* to *strongly agree*.

The aim of the proposal is to analyse the survey results via a sequential application of two approaches, non-parameteric and parametric, respectively. In particular, the results from one dimension of the first scale concerning the Learning Satisfaction Domain is synthesised into one ordinal response, defined via a joint data reduction (JDR) approach. To investigate the obtained ordinal outcome, accounting for possible heterogeneity that provides additional information on the effects of students' achievements, a cumulative model with proportional assumption and scale effect [10] has been considered. Furthermore, a recent recursive partitioning method yielding two trees, one for the location and one for the scaling, is also considered [12]. The method uses an algorithm which controls for the global significance level and selects the covariates having an impact on the ordinal response. The presentation is structured as follows: Section 2 briefly describes the generation process of the ordinal response whereas Section 3 illustrates the ordinal data model implemented for assessing the latent continuous variable (the DL perception). Some insights concludes the proposal.

## 2 Defining the ordinal outcome

In order to synthesise the students perspective on DL, we apply on the DL-related items a joint data reduction (JDR) approach. In the context of unsupervised learning, it is common practice to apply dimension reduction (e.g. principal component analysis, [7]) and then to cluster observations in the identified reduced space. The dimension reduction step is independent from the clustering step, and this may cause the two-step approach to misidentify the underlying structure of the data. JDR methods seek for a solution that is optimal for both steps: to this end, JDR methods consist of an iterative procedure that alternately optimise the data reduction and the clustering steps. Different JDR methods have been proposed, depending on the nature of the available attributes: JDR have been proposed for continuous, e.g., ([2],[15]), for categorical [5] and for mixed-type attributes (see [13] for a review). In this proposal  we refer to cluster correspondence analysis (cluster CA, [14]), a JDR method suitable for survey data. Let $\mathbf{Z}_j$ denote an $n \times p_j$ indicator matrix. That is, each row corresponds to a respondent, and the columns represent the $p_j$ levels of agreement for the $j^{th}$ item. Observed responses are coded by ones and all other elements are zero. The block matrix $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p]$ is the so-called super-indicator. The cluster CA procedure defines a cluster membership variable, that is obtained so that it minimises

$$\min \phi_{\text{CCA}}(\mathbf{B}^*, \mathbf{Z}_K) = \left\| \mathbf{D}_z^{-1/2} \mathbf{M} \mathbf{Z} - \mathbf{Z}_K \mathbf{G} \mathbf{B}^{*\prime} \right\|^2 \quad \text{s.t.} \quad \mathbf{B}^{*\prime} \mathbf{B}^* = \mathbf{I}_d \tag{1}$$

where $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n^{\prime}/n$ is a centering operator, $\mathbf{B}^* = \frac{1}{\sqrt{np}} \mathbf{D}_z^{1/2} \mathbf{B}$, $\mathbf{D}_z = diag\,(\mathbf{Z}'\mathbf{Z})$, $\mathbf{B}$ is the item weights matrix, and $\mathbf{Z}_K$ is the indicator version for the cluster membership variable.

The parameter $K$, that is user-defined, is set to four so that it matches the levels of the Likert scales. The JDR-based ordinal outcome is, therefore, the cluster membership variable, with levels sorted according to group characterisation. In particular, Figure 1 shows the items that characterise each group: each bar is the standardised residual from independence of the cross-table between the corresponding item and the cluster membership variable. The size of a bar is proportional to the negative or positive group characterisation made by the category corresponding to that bar. Since the items refer to statements describing a *favorable* DL perception, the groups are ranked according to the agreement towards the most characterizing items. For example, the cluster 2 (top-right plot in Figure 1) is characterised by *strong disagreement* towards DL and therefore is ranked as level 1 of the response outcome.

## 3 The location-scale model for Learning Satisfaction Domain

The ordinal response resulting from the JDR procedure is coded so that $Y_i$, with $i = 1, \ldots, n$, represents the grade expressed by the $i$-th student about the synthesis concerning the Learning Satisfaction Domain. For each $i$-th student, we also have $\boldsymbol{x}_i$, a row vector of the matrix $\boldsymbol{X}$ which includes all the students' characteristics and/or syntheses of the psychometric scales mentioned in the introduction. We indicate with $Y_i^*$ the underlying (continuous) latent variable related to Learning Satisfaction Domain such that, for any $i$-th subject,

$$\tau_{j-1} < Y_i^* \leq \tau_j \qquad \Longleftrightarrow \qquad Y_i = j\,, \qquad j = 1, 2, \ldots, K\,,$$

where $-\infty = \tau_0 < \tau_1 < \ldots < \tau_K = +\infty$ are the cut-points of $Y^*$.

Assuming that $p \geq 1$ covariates concerning characteristics of students or synthesis of the psychometric scales are relevant for explaining $Y^*$ by the latent regression model, we have

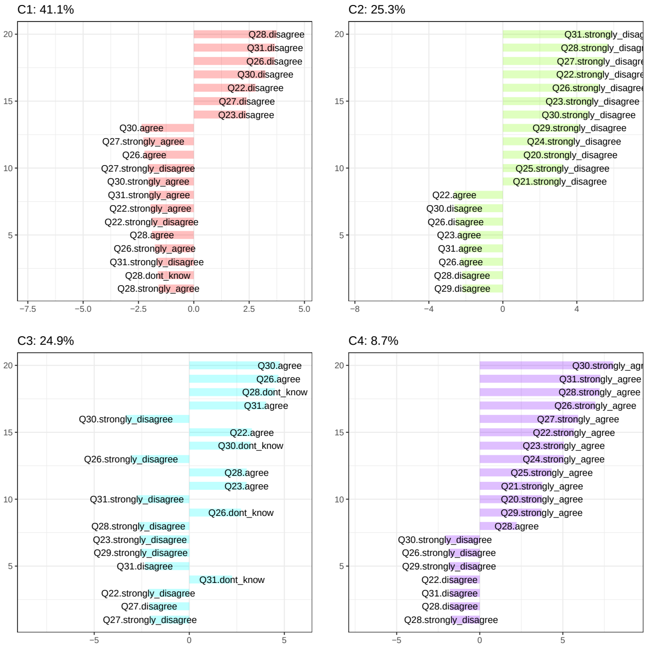

**Fig. 1** Item scores for groups characterisation: deviations from independence condition

$$Y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \sigma\varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

where $\sigma = \exp(\boldsymbol{z}_i\boldsymbol{\gamma})$ is the *relative* scale, $\boldsymbol{z}$ is an additional vector of covariates having impact on the scale and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)'$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_s)'$ the covariates coefficients. Then, the the location-scale model is:

$$Pr\,(Y_i j \mid \boldsymbol{\theta}, \boldsymbol{x}) = F_\varepsilon\left(\frac{\tau_j - \boldsymbol{x}_i\boldsymbol{\beta}}{\exp(\boldsymbol{z}_i\boldsymbol{\gamma})}\right).$$

Any strictly increasing distribution function may be conceived for $F_\varepsilon(.)$.

The model contains two terms that specify the impact of covariates, the location term $\tau_j - \boldsymbol{x}_i\boldsymbol{\beta}$ and the variance or scaling term $\exp(\boldsymbol{z}_i\boldsymbol{\gamma})$. If $\boldsymbol{x}$ and $\boldsymbol{z}$ are distinct the interpretation of the $\boldsymbol{x}$-covariates is the same as in the basic cumulative model with proportional assumption [10]. An alternative way to model heterogeneity, which has some advantages, specifies that covariates modify the cut-points. It has been discussed by [11].

In the proposal we concentrate our attention on the McCullagh's [10] location-scale model after having testing the presence of heterogeneity in the ordinal variable. The selection of relevant covariates for both components will be obtained by a recursive partitioning; a modeling strategy yielding a hybrid tree presented in [12].

Preliminary results show that students' age, resulted to be negative to Covid-19, studying experience, future employment, perception of the risk of infection, and social distancing affect the location of the Learning Satisfaction Domain whereas the fear of the risk of contagion impacts the scale of $Y^*$. Further analyses related to anxiety will be also discussed taking into account the connection between student stress and future career anxiety detected in the developments of a close research topic [6].

## References

1. Amir, L.R., Tanti, I., Maharani, D.A., Wimardhani, Y.S., Julia, V., Sulijaya, B., Puspitawati, R.: Student perspective of classroom and distance learning during covid-19 pandemic in the undergraduate dental study program universitas indonesia. BMC medical education **20**(1), 1–8 (2020)
2. De Soete, G., Carroll, J.D.: K-means clustering in a low-dimensional euclidean space. In: New approaches in classification and data analysis, pp. 212–219. Springer (1994)
3. Faraway, J.J., Augustin, N.H.: When small data beats big data. Statistics & Probability Letters **136**, 142–145 (2018)
4. Gray, E., Jennings, W., Farrall, S., Hay, C.: Small big data: Using multiple data-sets to explore unfolding social and economic change. Big Data & Society **2**(1), 2053951715589,418 (2015)
5. Hwang, H., Dillon, W.R., Takane, Y.: An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. Psychometrika **71**(1), 161–171 (2006)
6. Iannario, M., Iodice D'Enza, A., Romano, R.: Antecedents of distance learning perception of the students during the covid-19 pandemic. a partial least square - structural equation modeling. Manuscript (2021)
7. Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **374**(2065), 20150,202 (2016)
8. Lindstrom, M.: Small data: the tiny clues that uncover huge trends. St. Martin's Press (2016)
9. Mahmud, M.S., Talukder, M.U., Rahman, S.M.: Does 'fear of covid-19'trigger future career anxiety? an empirical investigation considering depression from covid-19 as a mediator. The International journal of social psychiatry (2020)
10. McCullagh, P.: Regression models for ordinal data. Journal of the Royal Statistical Society. Series B **42**, 109–142 (1980)
11. Tutz, G., Berger, M.: Separating location and dispersion in ordinal regression models. Econometrics and Statistics **2**, 131–148 (2017)
12. Tutz, G., Berger, M.: Tree-structured scale effects in binary and ordinal regression. Statistics and Computing **17**, 31–17 (2021). DOI https://doi.org/10.1007/s11222-020-09992-0
13. van de Velden, M., Iodice D'Enza, A., Markos, A.: Distance-based clustering of mixed data. Wiley Interdisciplinary Reviews: Computational Statistics **11**(3), e1456 (2019)
14. van de Velden, M., Iodice D'Enza, A., Palumbo, F.: Cluster correspondence analysis. Psychometrika **82**(1), 158–185 (2017)
15. Vichi, M., Kiers, H.A.: Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis **37**(1), 49–64 (2001)
16. Zurlo, M.C., Cattaneo Della Volta, M.F., Vallone, F.: Covid-19 student stress questionnaire: Development and validation of a questionnaire to evaluate students' stressors related to the coronavirus pandemic lockdown. Frontiers in Psychology **11**, 2892 (2020)