Journal of Classification
https://doi.org/10.1007/s00357-025-09521-6

ORIGINAL RESEARCH

Model-Based Clustering of Multivariate Rating Data Accounting for Feeling and Uncertainty

Matteo Ventura[1] · **Julien Jacques**[2] · **Paola Zuccolotto**[1]

Received: 17 December 2024 / Accepted: 10 October 2025
© The Author(s) 2025, modified publication 2026

Abstract
Among the models for the analysis of rating data, the CUB (combination of discrete uniform and shifted binomial random variable) is notable because it assumes that the final rating given by a respondent is the result of the joint action of two latent components: the feeling and the uncertainty, which are modelled through a shifted binomial and a uniform random variable, respectively. We propose a mixture of multivariate CUB models for clustering multivariate rating data, estimated using the EM algorithm. The performance of our model was evaluated through simulation studies, which demonstrated its capability to capture the hidden structure of the data. Additionally, we show two case studies which highlight the model's effectiveness in uncovering latent structures in real-world data and its utility in interpreting the results.

Keywords Ordinal data · Model-based clustering · EM algorithm · CUB class · Survey data

1 Introduction

Latent class models, also known as finite mixture models, play a crucial role in accounting for unobserved heterogeneity within groups of subjects exhibiting different latent behaviours. These models are instrumental in uncovering hidden structures within datasets across various research domains (Hagenaars & McCutcheon, 2002; Peel & MacLahlan, 2000) and find applications in cluster analysis by providing a flexible framework that can generalize classical geometric clustering methods (Bouveyron et al., 2019; McNicholas, 2016).

Clustering methods are widely used in several research areas for analyzing multivariate data and can handle different types of variables. In recent decades, research in several fields such as psychology, marketing, sociology, educational sciences, and behavioural sciences has focused its research efforts on measuring, analyzing, and studying latent constructs like perceptions, sensations, opinions, tastes, and attitudes related to specific topics or items. Information about these hidden traits is commonly obtained by administering questionnaires

✉ Matteo Ventura
matteo.ventura@unibs.it

[1] Department of Economics and Management, University of Brescia, Contrada S. Chiara, 50, 25122 Brescia, Italy

[2] ERIC Laboratory, Lumière University Lyon 2, 5 Av. Pierre Mendès, 69500 Bron, France

to people and asking the respondents to express their opinions. Questionnaires can be designed to collect different types of information through variables expressed on different response scales (Stevens, 1946).

Rating questionnaires are a common tool to investigate respondents' opinions, where individuals evaluate a given subject or item by selecting a score from a predefined set of ordered alternatives on a response scale. The resulting data are typically classified as ordinal categorical variables, which pose statistical challenges that require specialized methods and models tailored to accommodate their unique properties (Agresti, 2010, 2012).

Among the models for rating data proposed in the literature, the CUB (combination of a discrete uniform and shifted binomial random variable) class (Piccolo & Simone, 2019) is particularly intriguing. This class of models posits that the final ratings of the respondents are based on a combination of two latent components: the feeling and the uncertainty, modelled through a mixture distribution comprising a shifted binomial and a discrete uniform random variable, respectively (D'Elia & Piccolo, 2005; Piccolo, 2003). The CUB model is well-suited for rating data, especially those measured on the Likert scale, where responses are ordered based on increasing levels of satisfaction or agreement with the item (Likert, 1932).

Compared to other types of variables, the ordinal ones have received less attention, especially in the model-based clustering context. It is also common practice among researchers and practitioners to treat ordinal data as numerical, thereby ignoring their categorical nature and using models specifically designed for continuous data analysis, assuming more or less explicitly that the ratings derive from the discretization of an underlying latent continuous variable. Notable approaches include the binary ordinal search (BOS) algorithm (Biernacki & Jacques, 2016) and models assuming ordinal data as a generalization of latent Gaussian distributions (Grün & Dolnicar, 2016; McParland & Gormley, 2016; Ranalli & Rocci, 2017; Corneli et al., 2020). Another model-based approach considers logit models for taking into account the presence of covariates (Preedalikit et al., 2024).

Although there are numerous contributions and extensions to the CUB class even in a clustering context (Biasetton et al., 2024), to the best of our knowledge, a model-based clustering approach for multivariate data within this framework has not yet been proposed. Outside the CUB class, but with an approach that shares some similarities with it due to the presence of a uniform component, Browne et al. (2011) proposed a mixture of Gaussian and uniform distributions.

This article proposes a model-based approach that employs the CUB model to cluster multivariate rating data, using its assumptions and properties. Since the CUB model primarily addresses univariate data, we extend it to handle multivariate scenarios, resulting in a latent class extension of multivariate CUB models, abbreviated as MLC-CUB, tailored for clustering multivariate rating data.

The paper is organized as follows: in Section 2, the main characteristics of the CUB framework are explained, while in Section 3, the multivariate latent class CUB is defined, and the EM algorithm for the maximum likelihood estimation of the model parameters is derived. In Section 4, a simulation study is performed to assess the performance of the model, while in Section 4.2, the identifiability of the model is discussed, and a tool is proposed to detect the possible existence of multiple models fitting the same data. Subsequently, in Section 4.3, the model is compared to some mixture models commonly used for the analysis of ordinal data. After demonstrating the effectiveness of our model with simulated data, we show its application to real-world data in Section 5. Finally, the results and further developments are discussed in Section 6.

2 CUB Class of Models

The main idea of the CUB framework is that the reasoning of the respondents when faced with a Likert-type questionnaire can be modelled by a combination of two latent components, the feeling and the uncertainty, which respectively refer to the subjective experience of an individual towards an object or an event, and the inherent indecision present in every human choice that can be due, for example, to the lack of information or knowledge about the same object or event. The CUB class of models is founded on the basic CUB (D'Elia & Piccolo, 2005), which models the observed rating $r = 1, \ldots, m$, where m is the number of ordinal categories, as the realization of a mixed discrete random variable R with probability mass defined as follows:

$$P(R = r \mid \xi, \pi) = \pi P_B(r \mid \xi, m) + (1 - \pi) P_U(r \mid m), \tag{1}$$

where $\pi \in (0, 1]$ and $\xi \in [0, 1]$.

The distribution of the random variable R is a mixture of a shifted binomial probability distribution, P_B, and a discrete uniform probability distribution, P_U, defined over the support $\{1, \ldots, m\}$. Thus, the probability mass function of the basic CUB model (1) can also be written as follows:

$$P(R = r \mid \xi, \pi) = \pi \left[\binom{m-1}{r-1} \xi^{m-r} (1 - \xi)^{r-1} \right] + (1 - \pi) \frac{1}{m}, \tag{2}$$

where $r = 1, \ldots, m$ is the rating chosen by the respondent. The shifted binomial distribution accounts for the feeling, which is then measured by $1 - \xi$, and the uniform distribution for the uncertainty, measured by $1 - \pi$.

The log-likelihood function of CUB models is commonly optimized using the EM algorithm (D'Elia & Piccolo, 2005; Manisera & Zuccolotto, 2017; Piccolo & Simone, 2019; Simone, 2021; Manisera & Zuccolotto, 2022), which, albeit at the cost of slower computation, has proven to be more stable than numerical methods (Nelder & Mead, 1965), which are more sensitive to starting values and can experience convergence problems.

Among the CUB class, there are several notable models that fit the framework set by the basic CUB model. Some proposals include measuring different response styles and decision processes (Manisera & Zuccolotto, 2014b; Simone & Tutz, 2018; Manisera & Zuccolotto, 2022; Manisera et al., 2024), or extending the model to the multivariate case, including subject- and item-specific uncertainty (Simone et al., 2020; Colombi & Giordano, 2016; Andreis et al., 2013). Some researchers mixed the CUB model with well-known statistical methods and tools such as decision trees (Simone, 2023; Cappelli et al., 2019), clustering algorithms to cluster univariate CUB models (Corduas, 2008; Biasetton et al., 2023), and univariate latent class CUB models (Grilli et al., 2014).

3 Multivariate Latent Class CUB Model

In this section, the multivariate latent class CUB (MLC-CUB) model for clustering multivariate rating data is proposed. In the literature on CUB models, there are some proposals of multivariate CUB models (Simone et al., 2020; Colombi & Giordano, 2016; Andreis et al., 2013; Ip & Wu, 2024), but using these solutions for clustering purposes would lead to an overparametrised model. Therefore, we propose a more parsimonious model which simply consists of a mixture of multivariate CUB models, under the assumption of conditional independence, which is a common assumption for categorical data and which states that conditionally on the belonging to cluster k, the J ordinal responses of an individual

are independently drawn from J univariate CUB models. Let $\boldsymbol{R} = (R_j)_{j=1,\ldots,J}$ be a multivariate ordinal variable, where the jth component is an ordinal variable with m_j categories. Let ω_k be the mixing proportion of cluster k, such that $\omega_k > 0$ and $\sum_{k=1}^{K} \omega_k = 1$, and let $\boldsymbol{Z} : (Z_k)_{k=1,\ldots,K}$ be a group indicator variable distributed as a one-order multinomial distribution, $\boldsymbol{Z} \sim \mathcal{M}(1; \omega_1, \ldots, \omega_k)$, such that $Z_k = 1$ if the preferences come from cluster k, and $Z_k = 0$ otherwise. Given the conditional independence assumption, conditionally on cluster k and being $\boldsymbol{r} = (r_j)_{j=1,\ldots,J}$, assuming that $P_B(r_j \mid \xi_{jk}, m_j) = \binom{m_j-1}{r_j-1} \xi^{m_j-r_j}(1-\xi)^{r_j-1}$ and $P_U(r_j \mid m_j) = 1/m_j$, the distribution of $\boldsymbol{R}$ is defined as follows:

$$P(\boldsymbol{R} = \boldsymbol{r} \mid Z_k = 1, \boldsymbol{\pi}_k, \boldsymbol{\xi}_k) = \prod_{j=1}^{J} \Big[\pi_{jk} P_B(r_j \mid \xi_{jk}, m_j) + (1 - \pi_{jk}) P_U(r_j \mid m_j)\Big], \quad (3)$$

where $\boldsymbol{\xi}_k = (\xi_{jk})_{j=1,\ldots,J}$ and $\boldsymbol{\pi}_k = (\pi_{jk})_{j=1,\ldots,J}$.

Therefore, the marginal distribution of $\boldsymbol{R}$ is then as follows:

$$P(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\omega}) = \sum_{k=1}^{K} \omega_k \prod_{j=1}^{J} \Big[\pi_{jk} P_B(r_j \mid \xi_{jk}, m_j) + (1 - \pi_{jk}) P_U(r_j \mid m_j)\Big], \quad (4)$$

with $\boldsymbol{\xi} = (\xi_{jk})_{j=1,\ldots,J;k=1,\ldots,K}$, $\boldsymbol{\pi} = (\pi_{jk})_{j=1,\ldots,J;k=1,\ldots,K}$, $\boldsymbol{\omega} = (\omega_k)_{k=1,\ldots,K}$.

We employ a mixture of mixtures, a well-established method in model-based clustering, particularly successful for outlier detection through mixtures of contaminated Gaussians (Browne et al., 2011; Morris et al., 2019; Tomarchio et al., 2022). In our case, it is important to note that Grilli et al. (2014) pointed out that including a uniform component in the mixture of univariate CUB models can cause identification problems. To address this, they proposed keeping the parameter π constant across the different groups, and so the latent class CUB is defined as follows:

$$P(R = r) = \sum_{k=1}^{K} \pi \omega_k P_B(r \mid \xi_k, m) + (1 - \pi) P_U(r \mid m) \quad (5)$$

This approach is reasonable because in real applications, various factors influencing uncertainty might be common among subjects from different groups. Looking at Eq. (5), it is clear that the assumption proposed by Grilli et al. (2014) leads to the definition of a mixture of $K+1$ components, namely a uniform distribution and K shifted binomial distributions. Thus, since there is only one uniform component, the mixture is identifiable. However, applying this method to our case would mean assuming that $\pi_{jk} = \pi_j$. Due to the product over the J variables, the parametrization into $K + 1$ components reached in the univariate case is no longer valid. If we develop the product in Eq. (4), we obtain a mixture of $K \times 2^J$ terms. Among these terms, the term $1/m_j$ will appear K times, corresponding to $\prod_j P_U(r_j|m_j)$ which does not depend on k. Each of these terms being weighted with probability $\omega_k \prod_{j=1}^{J}(1 - \pi_{ik})$. This could lead to an identifiability problem, allowing to take any values for π_{ik} and ω_k such that $\sum_{k=1}^{K} \omega_k \prod_{j=1}^{J}(1 - \pi_{ik})$ is constant. One solution can be to collapse these terms into a single component, with probability $\sum_{k=1}^{K} \omega_k \prod_{j=1}^{J}(1 - \pi_{ik})$, leading to a model with $K \times 2^J - K + 1$ components. But the interpretation of such a model would be less easy than for the MLC-CUB model. Moreover, the probability that an observation comes from this component, $\sum_{k=1}^{K} \omega_k \prod_{j=1}^{J}(1 - \pi_{ik})$, is generally very low, in particular when J and/or π_{ik} are large. For these reasons, we keep the MLC-CUB model as defined by (4) and provide a heuristic diagnostic tool (Section 4.2) to identify whether, in practice, an identifiability problem can occur when estimating the model on a given dataset.

3.1 Maximum Likelihood Estimation with EM Algorithm

Let us consider an n-sample $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ where each $\boldsymbol{r}_i$ is an observation of the multivariate rating variable $\boldsymbol{R}$ which is independently and identically distributed, and drawn from an MLC-CUB distribution $P(\mathbf{R} \mid \boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\omega})$. In addition, let us consider $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ and $\boldsymbol{V}_i = (V_{i1}, \ldots, V_{iJ})$ which are the latent variables associated with each subject i. Since the parameters of finite mixture models cannot be estimated in a closed form, an effective way to overcome this issue is represented by the EM algorithm (Dempster et al., 1977), which allows us to obtain the maximum likelihood estimates of the model parameters by maximizing the complete log-likelihood function. This function is defined by introducing two latent allocation variables in the model. The random variable $\boldsymbol{Z} : (Z_k)_{k=1,\ldots,K}$ is distributed as a one-order multinomial distribution, $\boldsymbol{Z} \sim \mathcal{M}(1; \omega_1, \ldots, \omega_k)$, such that $Z_{ik} = 1$ if the ith rater preferences come from cluster k, and $Z_{ik} = 0$ otherwise. The second allocation variable $\boldsymbol{V} : (V_j)_{j=1,\ldots,J}$, is dependent on the allocation variable $\boldsymbol{Z}$, and it is distributed as a Bernoulli with parameter π_{jk}. The observation $V_{ij} = 1$ if the preference of the ith rater for the jth item comes from a shifted binomial of parameter ξ_{jk}, and $V_{ij} = 0$ if it belongs to a uniform random variable. Therefore, the complete log-likelihood is defined as follows:

$$\begin{aligned}\ell_c(\boldsymbol{\theta} \mid \mathbf{R}, \mathbf{Z}, \mathbf{V}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\Big\{ &\log(\omega_k) + \\ &+ \sum_{j=1}^{J} V_{ij}\big[\log(\pi_{jk}) + \log\big[P_B(R_{ij} \mid \xi_{jk}, m_j)\big]\big] + \\ &+ \sum_{j=1}^{J}(1 - V_{ij})\big[\log(1 - \pi_{jk}) + \log P_U(R_{ij} \mid m_j)\big]\Big\},\end{aligned} \tag{6}$$

where $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\omega})$.

Starting from the initial set of parameters generated at random, $\hat{\boldsymbol{\theta}}^{(0)} = (\hat{\boldsymbol{\xi}}^{(0)}, \hat{\boldsymbol{\pi}}^{(0)}, \hat{\boldsymbol{\omega}}^{(0)})$, the algorithm alternates between an expectation step and a maximization step.

At the t-th iteration, the expectation step consists of evaluating the expected value of the complete log-likelihood conditioned on the observed data $\boldsymbol{r}$ and on the parameters $\boldsymbol{\theta}^{(t)}$ computed in the previous step. Therefore, computing the conditional expectation of Eq. (6), for all $i = 1, \ldots, n, j = 1, \ldots, J, k = 1, \ldots K$, leads to compute the conditional expectation of Z_{ik} as follows:

$$\begin{aligned}\mathbb{E}(Z_{ik} \mid \boldsymbol{R}_i; \boldsymbol{\theta}^{(t)}) &= \frac{\omega_k^{(t)} \prod_j \left[\pi_{jk}^{(t)} P_B(R_{ij} \mid \xi_{jk}^{(t)}, m_j) + (1 - \pi_{jk}^{(t)}) P_U(R_{ij} \mid m_j)\right]}{\sum_{k'=1}^{K} \omega_{k'}^{(t)} \prod_j \left[\pi_{jk'}^{(t)} P_B(R_{ij} \mid \xi_{jk'}^{(t)}, m_j) + (1 - \pi_{jk'}^{(t)}) P_U(R_{ij} \mid m_j)\right]} = \\ &= \tau_{ik}(\boldsymbol{R}_i; \boldsymbol{\theta}^{(t)}) = \\ &= \tau_{ik}^{(t)},\end{aligned}$$

Then, the conditional expectation of the product $Z_{ik} V_{ij}$ is computed:

$$\begin{aligned}\mathbb{E}(Z_{ik} V_{ij} \mid R_{ij}; \boldsymbol{\theta}^{(t)}) &= \frac{\pi_{jk}^{(t)} P_B(R_{ij} \mid \xi_{jk}^{(t)}, m_j)}{\pi_{jk}^{(t)} P_B(R_{ij} \mid \xi_{jk}^{(t)}, m_j) + (1 - \pi_{jk}^{(t)}) P_U(R_{ij} \mid m_j)} \cdot \tau_{ik}(\mathbf{R}_i; \boldsymbol{\theta}^{(t)}) = \\ &= \nu_{ijk}(R_{ij}; \boldsymbol{\theta}^{(t)}) \cdot \tau_{ik}(\mathbf{R}_i; \boldsymbol{\theta}^{(t)}) = \\ &= \eta_{ijk}^{(t)},\end{aligned}$$

The maximization step, instead, consists of computing the new maximum likelihood estimates $\boldsymbol{\theta}^{(t+1)}$ of the parameters of the mixture, by maximizing according to $\boldsymbol{\theta}$ the function:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(t)} \Big\{ & \log(\omega_k) + \\ & + \sum_{j=1}^{J} \eta_{ijk}^{(t)} \big[\log(\pi_{jk}) + \log\big[P_B(R_{ij} \mid \xi_{jk}, m_j)\big]\big] + \\ & + \sum_{j=1}^{J} (1-\eta_{ijk}^{(t)}) \big[\log(1-\pi_{jk}) + \log P_U(R_{ij} \mid m_j)\big] \Big\}, \end{aligned} \tag{7}$$

By computing the first derivatives of the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$, the maximum likelihood estimators are obtained as follow:

$$\hat{\pi}_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n} \eta_{ijk}^{(t)}}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}, \quad \hat{\xi}_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n} \eta_{ijk}^{(t)} (m_j - r_{ij})}{\sum_{i=1}^{n} \eta_{ijk}^{(t)} (m_j - 1)}, \quad \hat{\omega}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)}}{n}. \tag{8}$$

The algorithm is stopped when a threshold $\epsilon > 0$ is reached in the relative change of the log-likelihood: $\mid \ell(\hat{\boldsymbol{\xi}}^{(t+1)}, \hat{\boldsymbol{\pi}}^{(t+1)}, \hat{\boldsymbol{\omega}}^{(t+1)}) - \ell(\hat{\boldsymbol{\xi}}^{(t)}, \hat{\boldsymbol{\pi}}^{(t)}, \hat{\boldsymbol{\omega}}^{(t)}) \mid < \epsilon$.

To avoid convergence to a local maximum, the algorithm is run multiple times from different random starting points. Among these iterations, the one yielding the highest log-likelihood is selected as the final outcome of the model.

3.2 Selection of the Number of Clusters

To determine the optimal number of clusters K, we employ the Bayesian information criterion (BIC; Schwarz, 1978). This criterion helps in selecting the model that best balances goodness of fit and model complexity.

The BIC for a number of clusters K is defined as follows:

$$\text{BIC}_K = -2 \log L(\boldsymbol{\theta} \mid \mathbf{r}) + (2KJ + K) \log(n),$$

where $L(\boldsymbol{\theta} \mid \mathbf{r})$ is the observed likelihood of the model, that is

$$L(\boldsymbol{\theta} \mid \mathbf{r}) = \prod_{i=1}^{n} \Big\{ \sum_{k=1}^{K} \omega_k \prod_{j=1}^{J} \Big[\pi_{jk} P_B(r_j \mid \xi_{jk}, m_j) + (1-\pi_{jk}) P_U(r_j \mid m_j) \Big] \Big\}.$$

To select the best number of clusters, the model has to be executed for all $K = 1, \ldots, \max_K$, where $\max_K$ is the maximum number of clusters we reasonably expect may be present in data, and the model with the lowest BIC_K is selected.

4 Simulation Studies and Performance Assessment

This section aims to evaluate through simulation studies the performance of the MLC-CUB model for clustering purposes regarding the influence of sample size and the robustness to noise. We also evaluated the efficiency of BIC in selecting the number of clusters. Furthermore, we investigated the model's identifiability in Section 4.2 and compared its performance

against established competitors in Section 4.3. All the simulations and applications of the model in this paper have been performed with R statistical software (R Core Team, 2024).

4.1 Simulation Design and Results

In all the experiments reported in this section, the EM algorithm is run ten times with uniform random starting parameters and with a threshold for the relative change of the log-likelihood equal to 10^{-5}. Among the ten runs, the one with the maximum log-likelihood has been kept.

The following three experiments were conducted by simulating 100 multivariate samples with sample sizes of $n = \{100, 500, 1000\}$. Each sample was generated from a trivariate ($J = 3$) mixture of two CUB components ($K = 2$), with each dimension having seven categories ($m_j = 7$ for all j). The parameters were selected to ensure low uncertainty since it is the case of most interest in practical applications, and the ξ parameters were sufficiently distinct for each variable. The mixture proportions were set to 0.25 and 0.75. The MLC-CUB parameters used for generating the datasets are presented in Table 1.

The computational time for running the prototype R code to estimate an MLC-CUB model is around 14 min for a 1000-observation dataset (using an AMD Ryzen 5 5600U with Radeon Graphics 2.30 GHz, 40GB RAM). This code will be optimized and included in the forthcoming R package. To facilitate immediate use, a temporary R script for fitting the model can be accessed at https://bodai.unibs.it/cub/mlc-cub/.

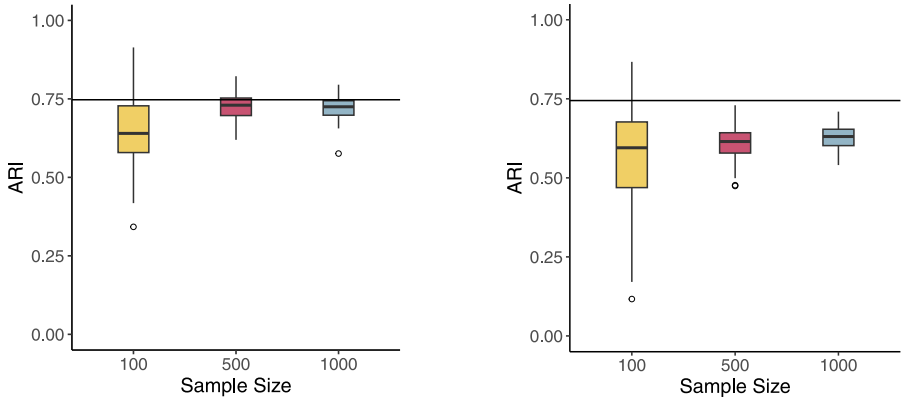
Influence of Sample Size This simulation study aims to investigate the influence of sample size on parameter estimates and the quality of the estimated partitions. This effect is studied both by looking at the distributions of the parameter estimates and at the quality of the estimated partitions which is assessed by using the adjusted Rand index (ARI; Hubert and Arabie, 1985). Values close to 1 indicate that the partition provided by the algorithm closely matches the simulated one, while values close to 0 suggest that the two partitions are no better aligned than random chance would predict. ARI results are shown in Fig. 1a, where an optimal ARI value (approximately $\simeq 0.75$) is indicated by a horizontal line. The optimal ARI has been computed by performing one step of the clustering algorithm with the true parameters on a simulated dataset with 10,000 observations and then comparing the obtained clustering with the real one (Jacques & Biernacki, 2014).

The boxplots show that increasing the sample size enhances the quality of clustering. For a sample size of 100, there is greater variability in the ARI. In contrast, for sample sizes of 500 and 1000, the ARI values are more consistent and closely aligned with the optimal ARI. The median ARI values for sample sizes equal to 500 and 1000 are almost equal to the optimal one. Those results suggest that larger sample sizes generally improve clustering accuracy; however, a medium sample size of 500 is still adequate to produce partitions of good quality.

In Figs. 2a, 3a, and 4a, the estimates of the parameters ξ_{jk}, π_{jk}, and ω_k are shown, respectively. These boxplots clearly illustrate that as the sample size increases, the variability

Table 1 Set of parameters chosen for generating the simulated datasets

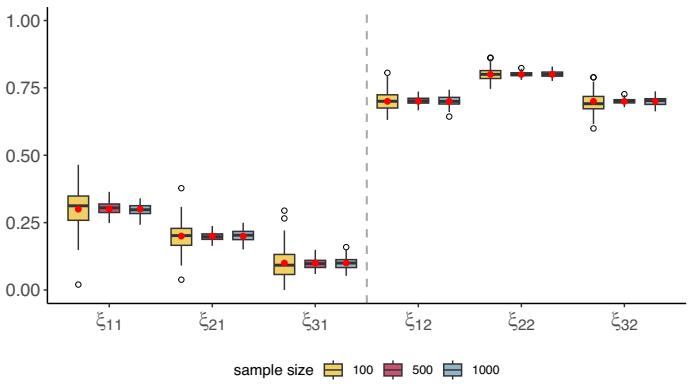
	$k = 1$			$k = 2$		
	$\omega = 0.25$			$\omega = 0.75$		
	$j = 1$	$j = 2$	$j = 3$	$j = 1$	$j = 2$	$j = 3$
π	0.80	0.90	0.60	0.60	0.80	0.70
ξ	0.30	0.20	0.10	0.70	0.80	0.70



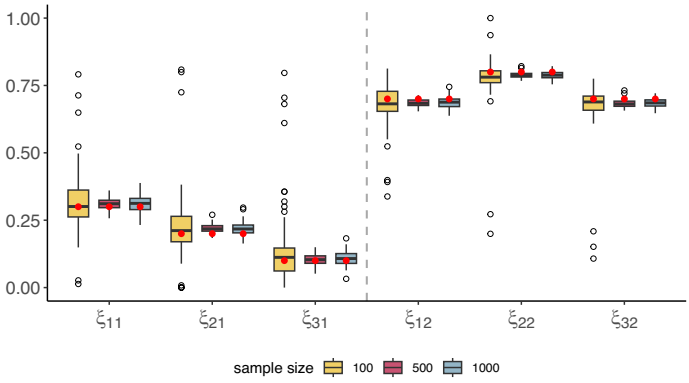
(a) Simulation 1 – Influence of sample size without noise.

(b) Simulation 2 – Influence of sample size with noisy data.

Fig. 1 Adjusted Rand index (ARI) distribution on 100 samples. The horizontal line represents the optimal ARI

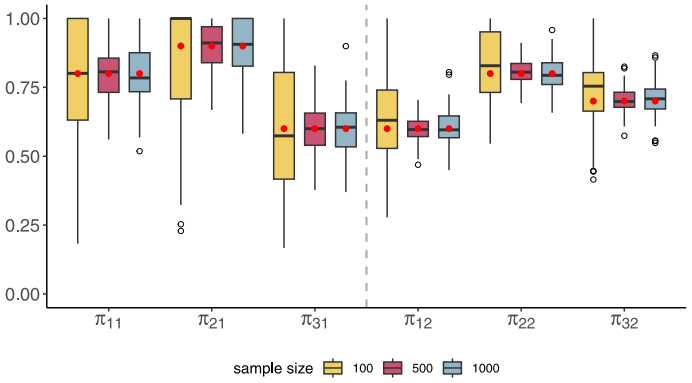


(a) Simulation 1 – effect of sample size without noise

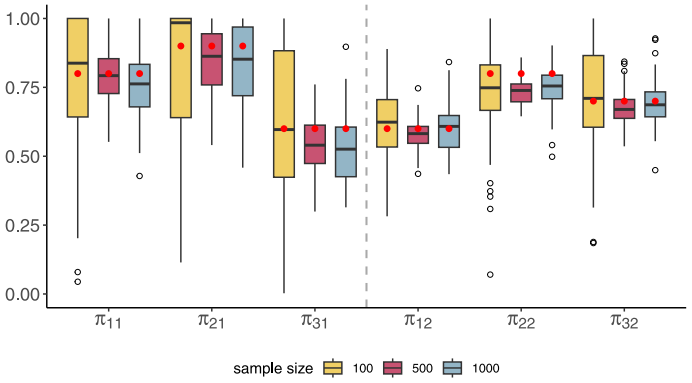


(b) Simulation 2 – effect of sample size with noisy data ($\phi = 0.1$)

Fig. 2 Boxplots representing the estimates of the parameter ξ_{jk} obtained on 100 datasets. The real values of the parameters are represented by the red circle

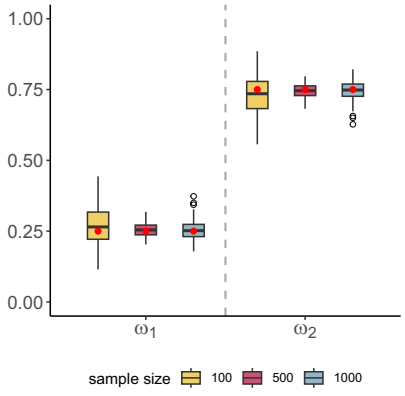


(a) Simulation 1 – effect of sample size without noise

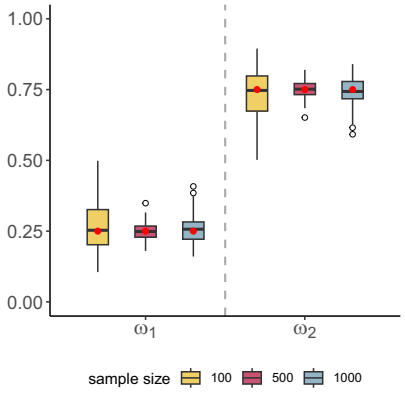


(b) Simulation 2 – effect of sample size with noisy data ($\phi = 0.1$)

Fig. 3 Boxplots representing the estimates of the parameter π_{jk} obtained on 100 datasets. The real values of the parameters are represented by the red circle



(a) Simulation 1 – effect of sample size without noise



(b) Simulation 2 – effect of sample size with noisy data ($\phi = 0.1$)

Fig. 4 Boxplots representing the estimates of the parameter ω_k obtained on 100 datasets. The real values of the parameters are represented by the red circle

of the estimates decreases. Noteworthy, with a medium and a large sample size, the median values of all parameter distributions coincide with the true parameter values, represented by red circles. It has to be noted that the estimates of the parameter π_{jk} with a small sample size are not so precise. Indeed, in this case, the median of the parameter estimates sometimes does not coincide with the real value of the estimate. It is important to highlight that the distributions of the parameters estimated on a small sample size, in our case $n = 100$, exhibit some outliers, indicating that with small sample sizes, the model encounters difficulties in accurately estimating the parameters. Instead, looking at the boxplots of the parameter estimates on datasets of size 500 and 1000, the distributions show lower variability and less outliers. This already happens with $n = 500$, confirming that a medium sample size is enough to obtain good parameter estimates.

Robustness to Noise This simulation aims to evaluate the performance and behaviour of the model when it is fitted on noisy data. Typically, noise is simulated by adding data from a uniform distribution to the dataset. However, in this specific case, introducing noise through a uniform random variable is not possible, as it is a component of the mixture that characterizes the CUB model since it models the uncertainty component. Therefore, we introduced noise in our simulated data by replacing some observations from a CUB component with observations from a discretized Gaussian distribution.

We simulated datasets with a noise proportion equal to 10% ($\phi = 0.1$). As in the previous simulations, the results were evaluated using the ARI (see Fig. 1b) and by examining the distributions of the parameter estimates (see Figs. 2b, 3b, and 4b). It is noteworthy that the variability of the ARI decreases with sample size, and the median value of the ARI approaches the optimal value as the sample size increases; however, in the presence of noise, the median value of ARI never reaches the optimal value. By looking at the distributions of the parameter estimates, it can be noticed that the parameter π_{jk} remains the most challenging to estimate, as demonstrated by its high variability. Interestingly, for the parameters ξ_{jk} and π_{jk}, even though the variability of the estimates decreases with increasing sample size, sometimes the median values of the parameter estimates do not coincide with the real ones, even though the deviation from the real value is very low. The only exception is the parameter ω_k, whose distribution is not much affected by the noise. Therefore, as expected, the quality of clustering and the accuracy of the parameter estimates decrease as the amount of noise increases, indicating that the model is sensitive to noise.

Selection of Number of Clusters This simulation study aims at evaluating the ability of our model to identify the correct number of clusters. To do that, we used the BIC, as described in Section 3.2, and a setting that consists of six different scenarios obtained by varying either the sample size $n \in \{100, 500, 1000\}$ or the amount of noise $\phi \in \{0, 0.1\}$. For each scenario, we draw 100 datasets. The results are reported in Table 2.

Table 2 Frequency of selection of the number of clusters K as the best number of clusters among 100 simulated datasets, for increasing number of observations and increasing amount of noise

	Scenario $\phi = 0$					Scenario $\phi = 0.1$				
n/K	1	2	3	4	5	1	2	3	4	5
100	18	82	-	-	-	56	44	-	-	-
500	-	100	-	-	-	-	100	-	-	-
1000	-	100	-	-	-	-	100	-	-	-

The maximum number of clusters that has been considered is $\max_K = 5$, and the actual number of clusters is $K = 2$

Looking at the results of the model selection, we can say that the performance of the model in selecting the correct number of clusters ($K = 2$) is overall good, especially when there is no noise. Indeed, it can be noticed that only in the case of a small sample size ($n = 100$) the model does not select the right number of clusters every time, both with and without noise. Indeed, in the case without noise, the model selects the right number of clusters in 82% of times, while this percentage decreases to 44% of times in the datasets with the addition of noise.

However, two facts have to be noticed: the first is that with a higher sample size, the model always selects the correct number of clusters; the second fact that has to be noticed is that a medium sample size ($n = 500$) is already enough to select the correct number of clusters in all the cases, both with and without noise.

4.2 Model Identifiability Assessment

As already discussed in Section 3, we decided to allow the π parameter to vary both across variables and clusters. This inevitably leads to identifiability issues due to the uniform component since it is equal for every cluster; therefore, when one observation r_{ij} comes from the uniform random variable (i.e. $v_{ij} = 0$), it is impossible to determine to which cluster it belongs (Titterington et al., 1985). However, we believe that, in the multivariate case, the probability of encountering this problem is lower than in the univariate case, since it should be enough that one observation r_{ij} comes from a shifted binomial to determine to which cluster it belongs. Additionally, we hypothesize that identifiability issues are less likely when the values of π_{jk} are high, indicating low uncertainty. This problem is extensive and complex; thus, it will be the subject of further research. Nevertheless, the identifiability problem also impacts real data analysis. Therefore, we propose an empirical procedure to detect identifiability problems in real datasets.

We guess that the distribution of the adjusted Rand index (ARI) computed between the partition obtained on the original sample and the partition obtained on each bootstrap sample is an effective tool for detecting models affected by identifiability issues. To illustrate our procedure, we generated three datasets with a sample size of $n = 1000$ to ensure that the sample size did not affect the estimation.

Dataset 1—No Identifiability Problem To observe the behaviour of an identifiable model, we generated a dataset identical to the one used in previous simulations, as we were reasonably confident that this model does not exhibit identifiability problems since the parameter estimates are good. This dataset consisted of three ordinal variables with seven categories, following an MLC-CUB distribution with parameters as shown in Table 1.

Dataset 2—Identifiability Problem with Two Clusters To investigate if the value of π_{jk} influences the identifiability of the model, we generated a dataset with three ordinal variables with seven categories, following a two-component MLC-CUB distribution characterized by low values of the parameter π_{jk}, with parameters as specified in Table 3.

Dataset 3—Identifiability Problem with Three Clusters To examine if the number of clusters affects identifiability problems, we generated a dataset with three ordinal variables with seven categories, following a three-component MLC-CUB distribution characterized by low values of the parameter π_{jk}, with parameters as specified in Table 4.

Dataset 4—Balanced Case To examine a more realistic case, we generated a dataset with two ordinal variables with seven categories, following a two-component MLC-CUB distri-

Table 3 Dataset 2—set of parameters chosen for generating the simulated data

	$k=1$			$k=2$		
	$\omega=0.25$			$\omega=0.75$		
	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
π	0.20	0.10	0.25	0.30	0.35	0.40
ξ	0.30	0.20	0.10	0.70	0.80	0.70

bution characterized by both low and high values of the parameter π_{jk}, with parameters as specified in Table 5.

We bootstrapped each dataset, and we estimated an MLC-CUB model for each one. Then, we computed the ARI between the partition obtained on the original sample and the partition obtained on each bootstrap sample.

The results of this procedure demonstrate that identifiability problems can be identified by examining the ARI distributions, which are shown in Fig. 1.

Indeed, when there are no identifiability problems, the distribution of the ARI is left-skewed and close to 1 (Fig. 5a), while when there are identifiability problems, the distribution of ARI is right-skewed and far from 1 (Fig. 5b and c). Figure 5d illustrates that in a more realistic scenario with mixed levels of uncertainty, the mode of the ARI distribution remains close to 1, despite increased variability. This suggests that the classification remains effective, albeit slightly less accurate than in the scenario depicted in Fig. 5a. However, as we will demonstrate in the case studies, real-world applications rarely exhibit extremely high levels of uncertainty. Therefore, this fourth scenario can be considered a worst-case scenario in practical contexts. Even under these conditions, the parameter estimates remain reliable and do not display bimodal distributions. Overall, the proposed model performs well even when uncertainty levels vary. Although this study does not establish whether the degree of skewness can measure the severity of identifiability problems, it does suggest that a left-skewed ARI distribution indicates stable results. Therefore, it is reasonable to infer that there are no identifiability problems in the model fitted to a specific dataset when the ARI distribution is left-skewed and has values close to 1. Notably, if we conduct the same experiment using a Gaussian mixture model (GMM; Banfield and Raftery, 1993), which is known for having no identifiability issue (if we omit label switching which has no impact on ARI), we obtained the distribution on Fig. 6a, which is very close to the one on Fig. 5a. Note that the GMM model has been estimated using the `Mclust` package (Scrucca et al., 2016). Such ARI distribution, instead, is due to the presence of clusters that overlap, as can be noticed by looking at Fig. 6b.

4.3 Comparison with State-of-the-Art Methods

Existing literature offers several model-based approaches for ordinal data, providing alternatives to our proposed model. Notably, the binary ordinal search model (BOS; Biernacki and Jacques, 2016) employs a binary search algorithm for ordinal data modelling and is imple-

Table 4 Dataset 3—set of parameters chosen for generating the simulated data

	$k=1$			$k=2$			$k=3$		
	$\omega=0.20$			$\omega=0.50$			$\omega=0.30$		
	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
π	0.10	0.10	0.35	0.20	0.25	0.10	0.15	0.17	0.22
ξ	0.50	0.20	0.10	0.70	0.80	0.70	0.10	0.60	0.90

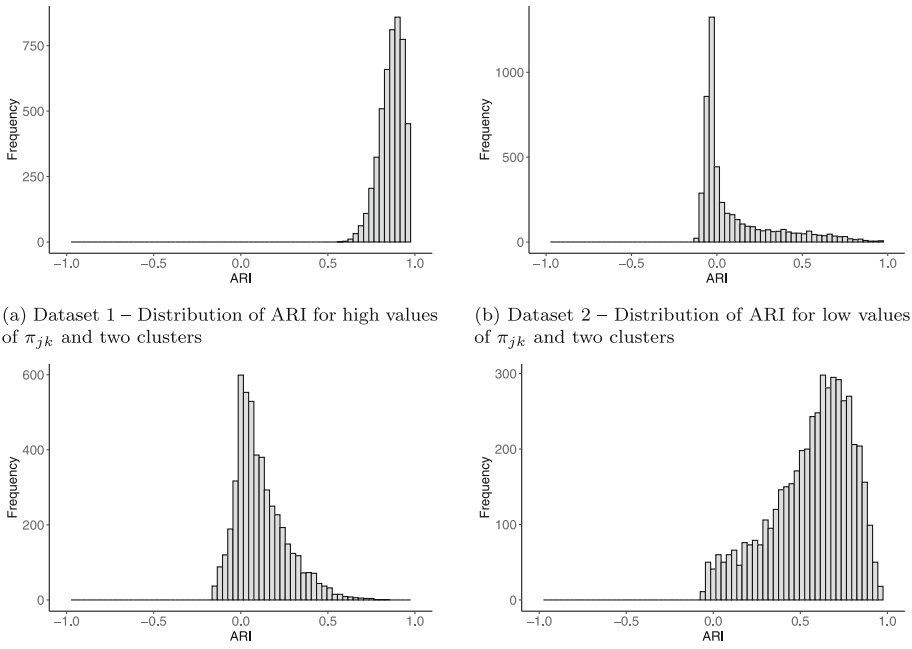
Table 5 Dataset 4—set of parameters chosen for generating the simulated data

	$k=1$			$k=2$		
	$\omega=0.25$			$\omega=0.75$		
	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
π	0.80	0.90	0.20	0.25	0.80	0.70
ξ	0.30	0.20	0.10	0.70	0.80	0.70

mented in the R package `ordinalClust` (Selosse et al., 2021). It is important to note that the BOS model's estimation performance is sensitive to the number of ordinal categories, demonstrating efficiency primarily for up to $m = 8$ categories (Biernacki & Jacques, 2016). Furthermore, it cannot handle variables with differing numbers of categories. Another relevant competitor is the clustMD model (McParland & Gormley, 2016), which, even though it is designed for clustering mixed data, can also be used for ordinal data by treating them as discretized manifestations of a continuous latent variable.

In our comparison study, we cannot consider the BOS model and the ClustMD model because they presented convergence problems with the type of data we generated.

In this comparative study, we evaluate the proposed MLC-CUB model against three alternative models: the ordinal latent block model (OLBM; Corneli et al., 2020), the multinomial mixture model (MMM; Lebret et al., 2015; Everitt, 1984), and the Gaussian mixture model (GMM; Banfield and Raftery, 1993). Notably, Corneli et al. (2020) adopted a latent variable approach similar to McParland and Gormley (2016), developing the OLBM for co-clustering



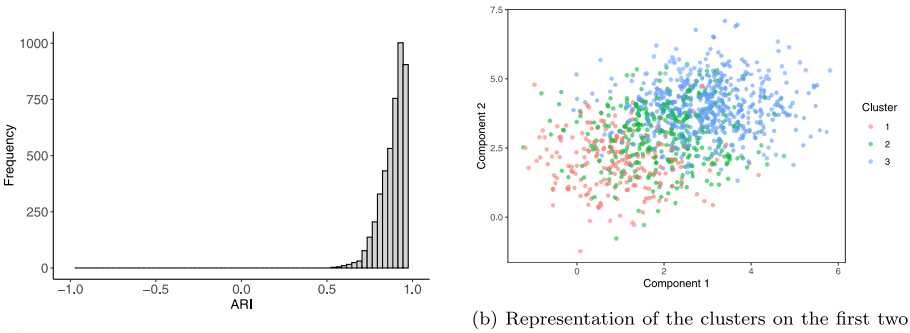
(a) Dataset 1 – Distribution of ARI for high values of π_{jk} and two clusters

(b) Dataset 2 – Distribution of ARI for low values of π_{jk} and two clusters

(c) Dataset 3 – Distribution of ARI for low values of π_{jk} and three clusters

(d) Dataset 4 – Distribution of ARI for mixed values of π_{jk} and two clusters

Fig. 5 Distribution of the ARI obtained as pairwise comparison of the classifications of 100 models fitted on three different bootstrapped datasets



(a) Distribution of ARI of the bootstrapped GMM

(b) Representation of the clusters on the first two dimensions

Fig. 6 Results of the clustering performed with GMM

ordinal data. While co-clustering simultaneously clusters rows and columns, we use OLBM in our context by constraining it to a single column cluster. The OLBM is implemented using the `OrdinalLBM` package (Corneli et al., 2020).

The MMM treats ordinal data as nominal by ignoring their ordinal nature and clustering them using mixtures of multivariate multinomials. It is fitted using the `Rmixmod` package (Lebret et al., 2015). In contrast, the GMM treats ordinal data as continuous, assuming they are generated from a mixture of multivariate Gaussian distributions. The GMM is fitted using the `Mclust` package (Scrucca et al., 2016).

To perform our comparison study, we generated 100 datasets with a sample size of $n = 1000$ with the same simulation setting described in Section 4; on each dataset, we fitted the MLC-CUB model, the OLBM, the GMM, and the MMM by keeping fixed the number of clusters ($K = 2$). We compared the clustering performance of each model by using the ARI, as shown in Fig. 7.

The results show that our model outperforms the others; the median ARI value approaches the optimal ARI. The ARI of the competing models is on average the same, but lower than the median ARI of MLC-CUB. Moreover, the ARI of OLBM and MMM shows high variability.

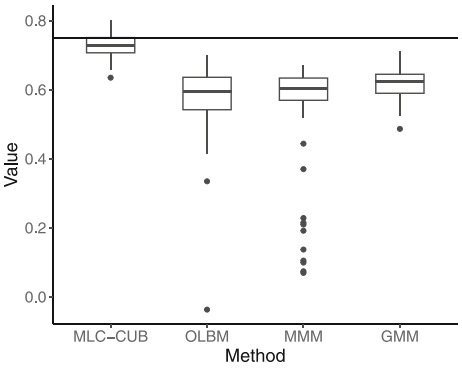


Fig. 7 ARI for multivariate latent class CUB model, ordinal latent block model, multinomial mixture model, and Gaussian mixture model. The horizontal line represents the optimal ARI

5 Case Studies

This section aims to validate the model's performance on real-world datasets through two illustrative case studies. By analyzing these practical applications, we can assess the model's strengths and its ability to extract useful information from multivariate data.

5.1 Evaluation of the University Orientation Service

The data analyzed in this section comes from the dataset `univer`, which is publicly available in the package `CUB` (Iannario et al., 2020). The dataset contains the answers from a sample survey on students' evaluation of the Orientation services of the University of Naples Federico II. The data were collected in 2002 and consisted of 2179 observations. Participants were asked to express their level of satisfaction on a seven-point Likert scale (1 = "very unsatisfied," 7 = "extremely satisfied") on the following elements of evaluation: (Q1) level of satisfaction about the acquired information; (Q2) level of satisfaction about the willingness of the staff; (Q3) level of satisfaction about the opening hours; (Q4) level of satisfaction about the competence of the staff; (Q5) level of global satisfaction.

We fitted the MLC-CUB model on the `univer` dataset, considering the five variables described above, and then we compared the results of our model with the competitor models by using the ARI and the BIC. The results are reported in Table 6.

The results show that both the MLC-CUB and MMM models identify five clusters in the data, whereas the GMM and OLBM models identify six clusters. Among all the models, the multinomial mixture model demonstrates the best performance in terms of BIC, having the lowest BIC value. The MLC-CUB model follows with the second-best BIC value.

The optimal MLC-CUB model is characterized by five components, with the estimated parameters for each cluster illustrated in Fig. 8. This plot is designed in line with the typical graphical representation of the basic CUB model, which is reported on a Cartesian plane that represents the parameter space of the CUB (D'Elia & Piccolo, 2005). In our representation, each univariate CUB model which constitutes the MLC-CUB model is plotted on the same Cartesian plane, where the x-axis reports the level of uncertainty and the y-axis reports the level of feeling.

By looking at the results, we can conclude that there are three main clusters (clusters 3, 4, 5), all characterized by low uncertainty and generally high levels of satisfaction. Among these three clusters, cluster 3 has the lowest level of satisfaction, although students in this group still exhibit medium-high satisfaction with university services.

Clusters 4 and 5 contain students with the highest levels of satisfaction. In particular, cluster 5 includes students who are completely satisfied with the university's orientation services, comprising 23% of the sample analysed.

Table 6 University dataset—values of the BIC for the MLC-CUB and its competitors for 1 to 7 clusters

Model	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$
MLC-CUB	33,891.10	30,253.74	28,769.16	28,301.50	**28,228.88**	28,251.68	28,289.33
OLBM	34,977.40	31,948.27	30,309.65	29,718.57	29,173.19	**28,676.43**	29,022.91
MMM	33,796.73	29,705.67	28,223.80	27,471.74	**27,292.52**	27,370.96	27,561.38
GMM	31,042.78	30,670.30	29,754.40	29,286.27	29,099.78	**29,069.44**	29,153.94

The best ones in row are highlighted in bold

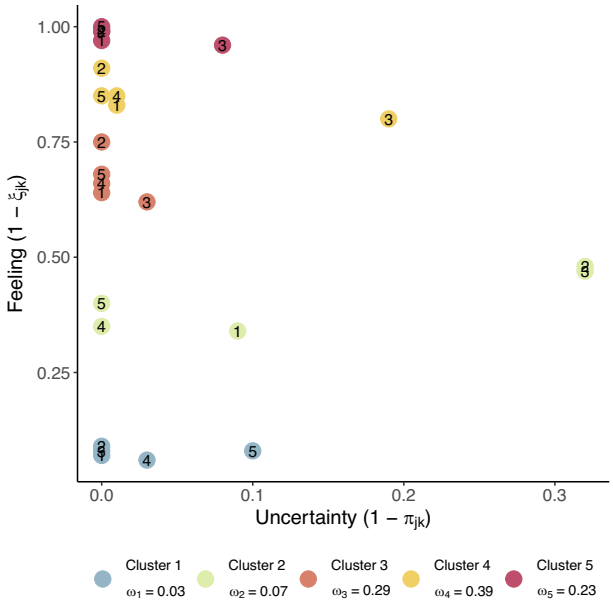


Fig. 8 University dataset—representation of the model parameters which characterize each cluster of the university data. Each number represents the associated question in the questionnaire

Clusters 1 and 2 are smaller compared to the others (see the legend of Fig. 8), but they are worth noting. Cluster 1 includes students who are not satisfied at all, while cluster 2 consists of students with medium-low levels of satisfaction, indicating that they are not completely satisfied with the university's orientation services. Additionally, students in cluster 2 exhibit higher levels of uncertainty in their responses to questions 2 (willingness) and 3 (office hours).

To assess the stability of the results and whether there were identifiability issues, the procedure described in Section 4.2 has been performed. The distribution of the pairwise ARI between the classifications fitted on 100 bootstrapped datasets is reported in Fig. 9.

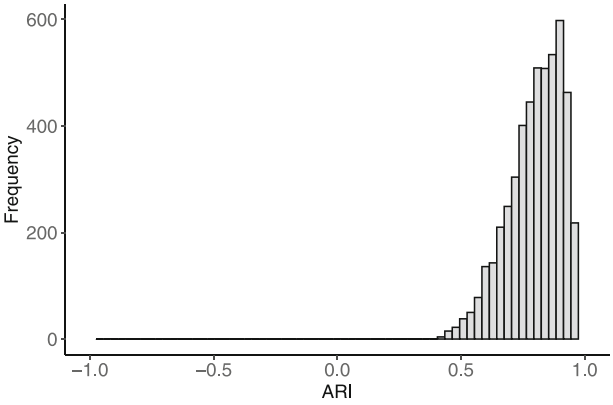


Fig. 9 University dataset—distribution of the pairwise ARI indexes between the classification fitted on bootstrapped data

The shape of the distribution is closely similar to the one reported in Fig. 5a, suggesting that there are no identifiability problems in the estimated model.

5.2 Evaluation of the Services in Public Kindergartens

In this subsection, we show an application of our model to real data with a focus on the evaluation of the services in public kindergartens (Brentari et al., 2006). The data were collected during a study sponsored by the Municipality of Brescia, Italy, in 2004. The aim of the survey was to evaluate the quality of service provided by the 22 municipal kindergartens. The questionnaire, divided into sections, consisted of questions where respondents expressed their agreement or disagreement on a four-point Likert scale, where 1 corresponded to "strongly disagree" and 4 corresponded to "strongly agree." There was also a fifth option, "Don't know" (DK), for respondents who were unsure of their judgment. The aspects on which parents were asked to provide their opinions included key characteristics of the service offered by the schools: the educational area, the relational area, and the organizational area. A total of 1337 responses were obtained. After the deletion of the observations that contained a DK answer or missing values, we obtained a dataset with 883 observations.

We decided to focus on the evaluation of the organizational area of the schools, assessed through the following questions: (Q1) Do you think the school environment is suitable for the activities conducted there? (Q2) Do you find the role of school aides (parents, cooks, etc.) to be adequate? (Q3) Are you satisfied with the quality and variety offered by the school cafeteria? (Q4) Are the reception hours for children adequate for your needs? (Q5) Do you consider the management of school entry and exit adequate? (Q6) Are you satisfied with the level of supervision provided by the school for the children? (Q7) Are you interested in the initiatives promoted for parents? And (Q8) Do you find parental participation in school activities adequate?

The obtained results have been compared with those obtained using other state-of-the-art methods described in Section 4.3. The optimal number of clusters for each method was determined using the BIC. The results are reported in Table 7. Three methods suggest that the best number of clusters is 3. However, one of those clusters is very small and does not add any valuable information to the interpretation of the results. Moreover, the difference in BIC between two and three clusters is low. Therefore, to enhance the interpretability of the results, we chose to present the model with two clusters.

The results of MLC-CUB are reported in Fig. 10. Looking at the results, we can conclude that there are two clusters of almost the same size (see the legend of Fig. 10). The first one is characterized by a higher level of feeling and, therefore, satisfaction with respect to the organizational level of the kindergarten system in Brescia. Both clusters show slightly

Table 7 Kindergarten dataset—values of the BIC for the MLC-CUB and its competitors for 1 to 7 clusters

Model	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$
MLC-CUB	13,350.31	12,767.05	**12,766.37**	12,809.55	12,888.92	12,981.69	13,070.47
OLBM	14,084.18	13,795.02	13,786.37	**13,773.98**	13,859.51	13,913.55	13,909.78
MMM	13,371.38	12,714.84	**12,693.21**	12,741.10	12,817.28	12,909.22	13,031.63
GMM	14,206.66	13,588.47	**12,526.45**	12,641.75	12,065.22	11,620.49	10,466.07

The best ones in row are highlighted in bold

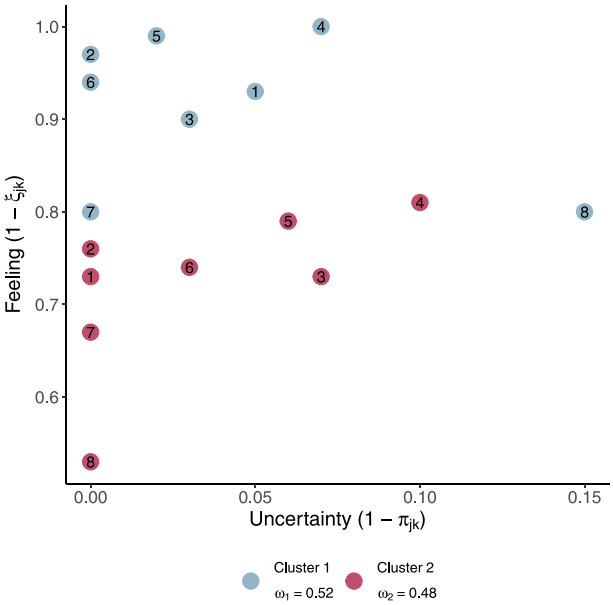


Fig. 10 Kindergarten dataset—representation of the model parameters which characterize each cluster of the kindergarten data. Each number represents the associated question in the questionnaire

lower levels of satisfaction in response to questions 7 and 8, which pertain to the level of parental participation and the parent-school relationship. In particular, both clusters exhibit a relatively low level of uncertainty.

Also in this application, we assessed the stability of the results by fitting the model on 100 bootstrapped datasets. We checked the identifiability of the estimated model by looking at the distribution of the pairwise ARI (see Fig. 11). The distribution has a shape similar to that reported in Fig. 5a, suggesting that the parameter estimates are stable and the estimated model is not compromised by identifiability problems.

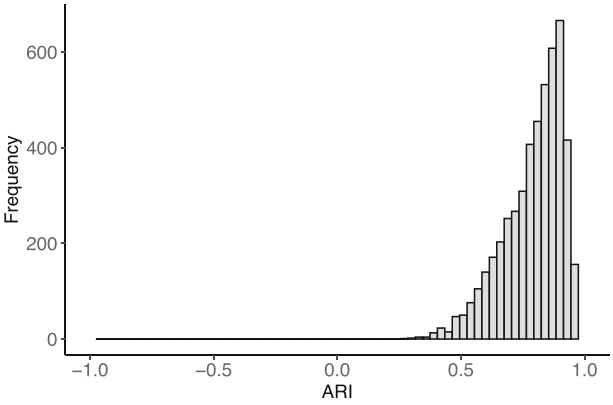


Fig. 11 Kindergarten dataset—distribution of the pairwise ARI indexes between the classification fitted on boostrapped data

6 Discussion

We presented a mixture model for clustering multivariate rating data, which belongs to the CUB class of models. Under the assumption of conditional independence, our model is a mixture of multivariate CUB models. This allows us to capture the latent components that characterize the CUB framework: feeling and uncertainty.

An EM algorithm for the maximum likelihood estimation of the model parameters has been defined and implemented. The performance of the algorithm was evaluated through simulation studies. The model demonstrated good performance when the level of uncertainty was low (i.e. the weight of the uniform component was low). However, it exhibited some drawbacks related to identifiability issues when the uncertainty was high. We explored this problem and proposed an empirical tool for identifying it. However, this issue will be the focus of further and more in-depth studies to better understand the conditions under which this problem arises and to determine which types of constraints on model parameters could be useful for ensuring identifiability.

Applications to real data demonstrated that the identifiability problem does not affect the model's performance in practice. In addition, these applications highlighted the valuable interpretability of the results in terms of understanding the latent traits that influence how people respond to questionnaires.

In the future, we plan to enhance the performance of the model by implementing more efficient initialization strategies, using, for instance, small EM algorithm, stochastic EM, or classification EM (Biernacki, 2004; Baudry & Celeux, 2015). Additionally, we plan to extend the model by incorporating the possibility to handle missing data imputation within the EM algorithm. Additionally, we aim to address the "Don't know" option in rating data, following the approach proposed by Manisera and Zuccolotto (2014a).

We also plan to extend the model in various ways. This includes incorporating other models within the CUB class to capture specific aspects and traits of response styles, such as using the discretized beta distribution (Simone & Tutz, 2018). Additionally, in this work, we assumed independence among ordinal variables, but we aim to relax this assumption using copulas. Another extension involves developing a multilevel latent class model for clustering multilevel rating data. Furthermore, a dedicated R package for this model is currently under development.

Acknowledgements We would like to thank Prof. Rosaria Simone for the valuable discussions and helpful advice regarding the CUB class of models. We are also grateful to the associate editor and the anonymous reviewers whose comments and suggestions greatly improved the quality of this manuscript.

Author Contributions Conceptualization: Julien Jacques, Matteo Ventura Methodology: Julien Jacques, Matteo Ventura Formal analysis and investigation: Matteo Ventura Writing - original draft preparation: Matteo Ventura Writing - review and editing: Julien Jacques, Matteo Ventura, Paola Zuccolotto Funding acquisition: Paola Zuccolotto Supervision: Julien Jacques, Paola Zuccolotto.

Funding Open access funding provided by Università degli Studi di Brescia within the CRUI-CARE Agreement. Part of this study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1551.11-10-2022, PE00000004 CUP D73C22001250001). This manuscript reflects only the authors' views and opinions; neither the European Union nor the European Commission can be considered responsible for them.

Data Availability The empirical data include two sources. The first, the `univer` dataset, is publicly available to all researchers. It can be easily accessed and downloaded via the popular R package `CUB`, available on CRAN. The second, the Kindergarten Dataset, is not publicly available due to privacy and usage restrictions

associated with the original collection. Interested researchers may obtain access to this data by submitting a formal request to the authors of the original cited paper that developed and introduced the dataset.

Declarations

Competing Interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Agresti, A. (2010). *Analysis of ordinal categorical data* (Vol. 656). John Wiley & Sons.

Agresti, A. (2012). *Categorical data analysis* (Vol. 792). John Wiley & Sons.

Andreis, F., Ferrari, P. A., et al. (2013). A proposal for the multidimensional extension of CUB models. In: *Cladag 2013: 9th meeting of the classification and data analysis group: book of abstracts* (pp. 15–18). CLEUP.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics,* 803–821.

Baudry, J.-P., & Celeux, G. (2015). EM for mixtures: Initialization requires special care. *Statistics and Computing, 25*, 713–726.

Biasetton, N., D'Urso, P., Disegna, M., & Salmaso, L. (2024). Cub model-based clustering of Likert-type data with a tourist satisfaction application. *Annals of Operations Research,* 1–16.

Biasetton, N., Disegna, M., Barzizza, E., & Salmaso, L. (2023). A new adaptive membership function with CUB uncertainty with application to cluster analysis of Likert-type data. *Expert Systems with Applications, 213*, 118893.

Biernacki, C. (2004). Initializing EM using the properties of its trajectories in Gaussian mixtures. *Statistics and Computing, 14*, 267–279.

Biernacki, C., & Jacques, J. (2016). Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing, 26*, 929–943.

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science: With applications in R* (Vol. 50). Cambridge University Press.

Brentari, E., Carpita, M., & Zuccolotto, P. (2006). *Qualità e customer satisfaction nei servizi: un'indagine statistica nelle scuole dell'infanzia del Comune di Brescia* (Vol. 472). FrancoAngeli.

Browne, R. P., McNicholas, P. D., & Sparling, M. D. (2011). Model-based learning using a mixture of mixtures of Gaussian and uniform distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(4), 814–817.

Cappelli, C., Simone, R., & Di Iorio, F. (2019). CUBREMOT: A tool for building model-based trees for ordinal responses. *Expert Systems with Applications, 124*, 39–49.

Colombi, R., & Giordano, S. (2016). A class of mixture models for multidimensional ordinal data. *Statistical Modelling, 16*(4), 322–340.

Corduas, M. (2008). A statistical procedure for clustering ordinal data. *Quaderni di Statistica, 10*, 177–189.

Corneli, M., Bouveyron, C., & Latouche, P. (2020). Co-clustering of ordinal data via latent continuous random variables and not missing at random entries. *Journal of Computational and Graphical Statistics, 29*(4), 771–785.

D'Elia, A., & Piccolo, D. (2005). A mixture model for preferences data analysis. *Computational Statistics & Data Analysis, 49*(3), 917–934.

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), 1–22.
Everitt, B. S. (1984). *Introduction to latent variable models*.
Grilli, L., Iannario, M., Piccolo, D., & Rampichini, C. (2014). Latent class CUB models. *Advances in Data Analysis and Classification, 8*, 105–119.
Grün, B., & Dolnicar, S. (2016). Response style corrected market segmentation for ordinal data. *Marketing Letters, 27*, 729–741.
Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.
Hubert, L., & Arabie, P. (1985). *Comparing partitions. Journal of classification, 2*, 193–218.
Iannario, M., Piccolo, D., & Simone, R. (2020). *The R package CUB: A class of mixture models for ordinal rating data*.
Ip, R. H., & Wu, K. Y. K. (2024). A mixture distribution for modelling bivariate ordinal data. *Statistical Papers,* 1–36.
Jacques, J., & Biernacki, C. (2014). Model-based clustering for multivariate partial ranking data. *Journal of Statistical Planning and Inference, 149*, 201–217.
Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., & Govaert, G. (2015). Rmixmod: The R package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library. *Journal of Statistical Software, 67*, 1–29.
Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology, 22*(140), 5–55.
Manisera, M., Migliorati, M., Ventura, M., & Zuccolotto, P. (2024). A mixture model for the analysis of categorical variables measured on five-point semantic differential scales. *Austrian Journal of Statistics, 53*(3), 70–86.
Manisera, M., & Zuccolotto, P. (2014). Modeling "Don't know" responses in rating scales. *Pattern Recognition Letters, 45*, 226–234.
Manisera, M., & Zuccolotto, P. (2014). Modeling rating data with nonlinear CUB models. *Computational Statistics & Data Analysis, 78*, 100–118.
Manisera, M., & Zuccolotto, P. (2017). Estimation of nonlinear CUB models via numerical optimization and EM algorithm. *Communications in Statistics-Simulation and Computation, 46*(7), 5723–5739.
Manisera, M., & Zuccolotto, P. (2022). A mixture model for ordinal variables measured on semantic differential scales. *Econometrics and Statistics, 22*, 98–123.
McNicholas, P. D. (2016). *Mixture model-based classification*. Chapman and Hall/CRC.
McParland, D., & Gormley, I. C. (2016). Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification, 10*(2), 155–169.
Morris, K., Punzo, A., McNicholas, P. D., & Browne, R. P. (2019). Asymmetric clusters and outliers: Mixtures of multivariate contaminated shifted asymmetric laplace distributions. *Computational Statistics & Data Analysis, 132*, 145–166.
Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308–313.
Peel, D., & MacLahlan, G. (2000). *Finite mixture models*. John & Sons.
Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica, 5*(1), 85–104.
Piccolo, D., & Simone, R. (2019). The class of CUB models: statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications, 28*, 389–435.
Preedalikit, K., Fernández, D., Liu, I., McMillan, L., Nai Ruscone, M., & Costilla, R. (2024). Row mixture-based clustering with covariates for ordinal responses. *Computational Statistics, 39*(5), 2511–2555.
Ranalli, M., & Rocci, R. (2017). A model-based approach to simultaneous clustering and dimensional reduction of ordinal data. *Psychometrika, 82*, 1007–1034.
Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics,* 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), 289.
Selosse, M., Jacques, J., & Biernacki, C. (2021). ordinalClust: An R Package to Analyze Ordinal Data. *The R Journal, 12*(2), 173–188.
Simone, R. (2023). Uncertainty Diagnostics of Binomial Regression Trees for Ordered Rating Data. *Journal of Classification,* 1–27.
Simone, R. (2021). An accelerated EM algorithm for mixture models with uncertainty for rating data. *Computational Statistics, 36*(1), 691–714.
Simone, R., & Tutz, G. (2018). Modelling uncertainty and response styles in ordinal data. *Statistica Neerlandica, 72*(3), 224–245.
Simone, R., Tutz, G., & Iannario, M. (2020). Subjective heterogeneity in response attitude for multivariate ordinal outcomes. *Econometrics and Statistics, 14*, 145–158.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680.
Titterington, D. M., Smith, A. F., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley.
Tomarchio, S. D., Gallaugher, M. P., Punzo, A., & McNicholas, P. D. (2022). Mixtures of matrix-variate contaminated normal distributions. *Journal of Computational and Graphical Statistics, 31*(2), 413–421.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.