A Mixture of Multivariate CUB Models for Clustering Rating Data

Matteo Ventura[1], Julien Jacques[2], Paola Zuccolotto[1]

[1] Department of Economics and Management, University of Brescia, Italy
[2] ERIC Laboratory, University Lumière Lyon 2, France

The idea

We propose a **Mixture Model** to cluster **rating data** derived from **Likert scales**.

- Likert scales are commonly used in questionnaires to measure respondents' opinions.
- One of the most notable models for analyzing such data is the CUB model.

The CUB Framework

Assumption: the **underlying Decision Process** leading to respondents' final ratings is characterized by two latent components:

Feeling: Reasoned and logical thinking, the set of emotions that individuals have with regard to the latent trait being evaluated.

- Modeled by a shifted Binomial:

$$P_B(\xi) = \binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}$$

- Measured by the **feeling parameter** $1-\xi$.

Uncertainty: Indecision inherently present in any human choice.

- Modeled by a discrete Uniform:

$$P_U(m) = \frac{1}{m}$$

- Measured by the **uncertainty parameter** $1-\pi$.

The final distribution is obtained as a **Combination of Uniform and shifted Binomial** [D'Elia and Piccolo, 2005], the **CUB** model:

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

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

The MLC-CUB model

To cluster multivariate rating data $\boldsymbol{R}$ with J independently and identically distributed ordinal variables, we propose the **Multivariate Latent Class CUB** (MLC-CUB) model:

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

with K being the number of clusters, $\boldsymbol{\pi} = (\pi_{jk})$, $\boldsymbol{\xi} = (\xi_{jk})$, $\boldsymbol{\omega} = (\omega_k)$ for $k = 1, \dots, K$ and $j = 1, \dots, J$.

- Estimation via **EM algorithm**.
- Uncertainty and feeling vary both across clusters and variables.
- It is possible to manage different numbers of categories.

Simulation study

The performances of our model have been **compared** with:

- Ordinal Latent Block Model (OLBM) [Corneli et al., 2020]
- Gaussian Mixture Model (GMM)
- Multinomial Mixture Model (MMM)

To study the **effect of sample size**, 100 data sets with sample size $n \in \{100, 500, 1000\}$ have been simulated from an MLC-CUB model with the following parameters:

	$k=1$			$k=2$		
ω	0.25			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

Table 1: Parameters set

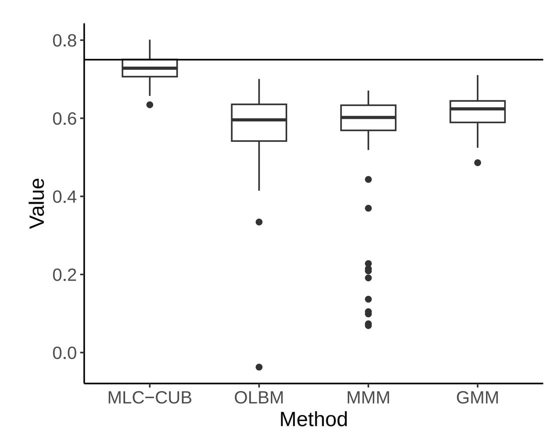


Figure 1: ARI for each compared model. The horizontal line represents the optimal ARI.

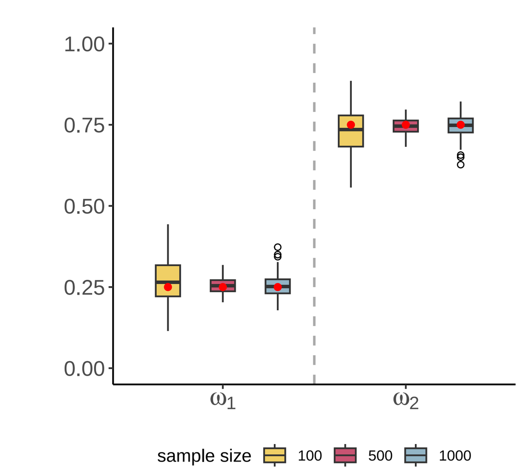


Figure 2: Effect of sample size on the estimates of the parameter ω_k.

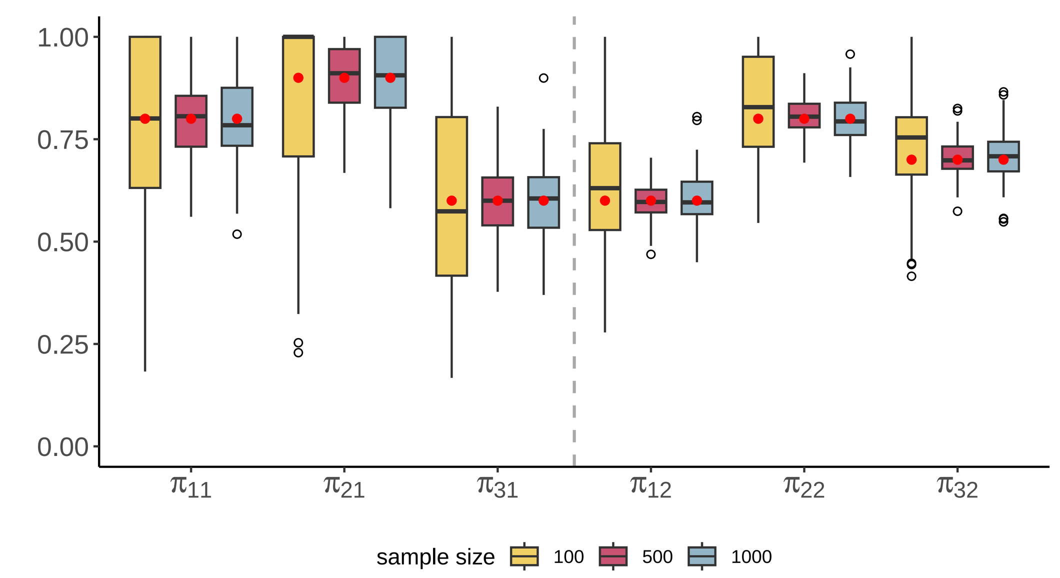


Figure 3: Effect of sample size on the estimates of the parameter π_{jk}.

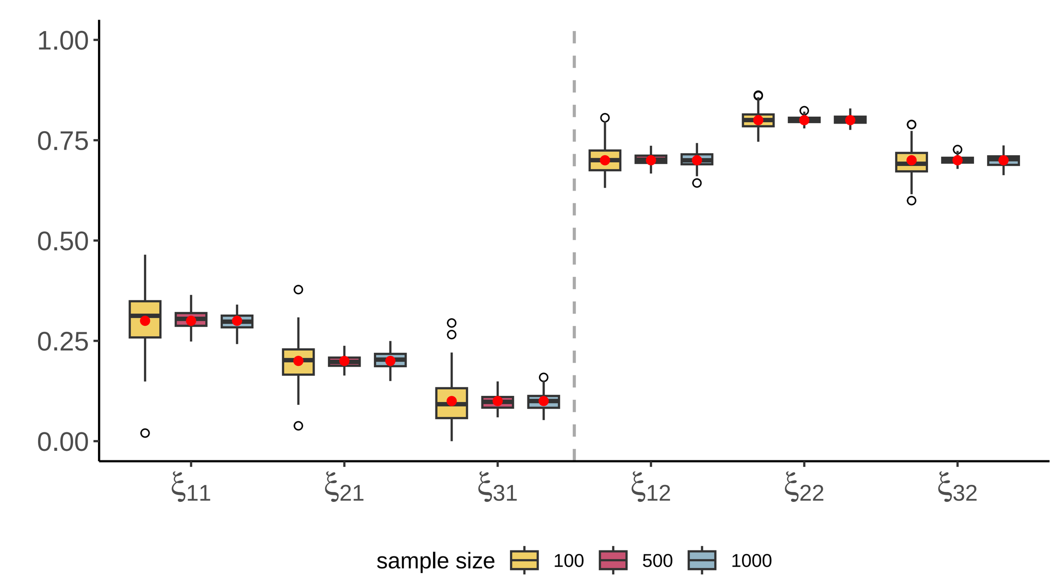


Figure 4: Effect of sample size on the estimates of the parameter ξ_{jk}.

Case study

Evaluation of the University Orientation Service

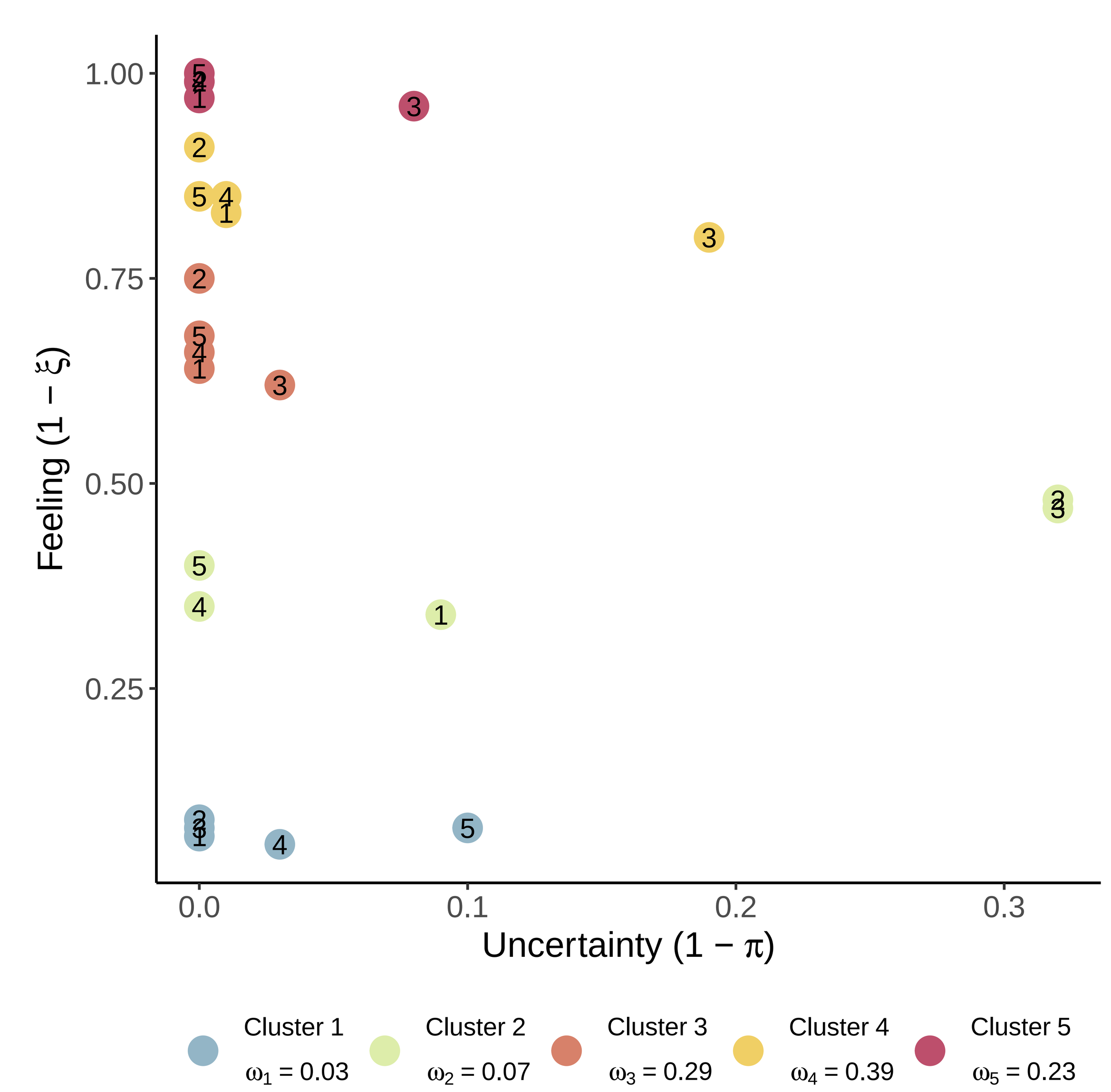
Data: `univer` data set (publicly available in the R package `CUB`).
Collection: sample survey.
Aim: evaluating the **students' satisfaction** about the Orientation services of the University of Naples Federico II, Italy.
Variables: five ($J = 5$) different aspects were evaluated:

1. Acquired information
2. Willingness of the staff
3. Opening hours
4. Competence of the staff
5. Global satisfaction

Total observations: 2179

Interpretation

- Three main clusters (clusters 3, 4, 5) characterized by low uncertainty and generally high levels of satisfaction.
- Two minor clusters:
  - Cluster 1 includes students who are not satisfied at all.
  - Cluster 2 includes students with a medium-low level of satisfaction.



Identifiability

The model is **not identifiable** due to the **Uniform** component.
Preliminary study of identifiability: simulation of 100 data sets with sample size $n = 1000$ from two MLC-CUB models characterized by:

- low values of the parameters π_{jk} (**high uncertainty**);
- high values of the parameters π_{jk} (**low uncertainty**).

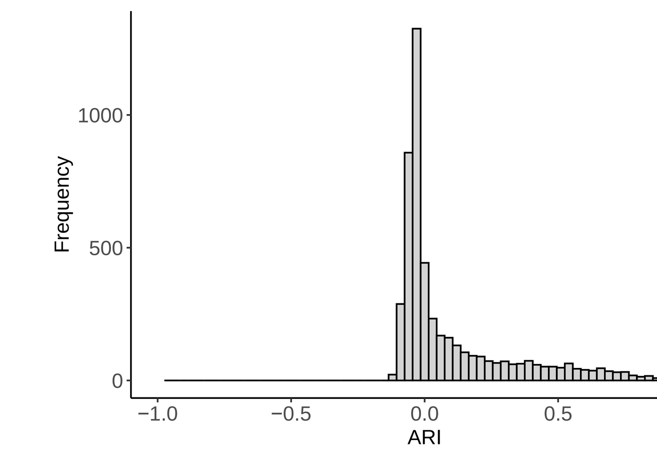


Figure 5: **Ientifiability problem** – Distribution of ARI when the values of π_{jk} are low.

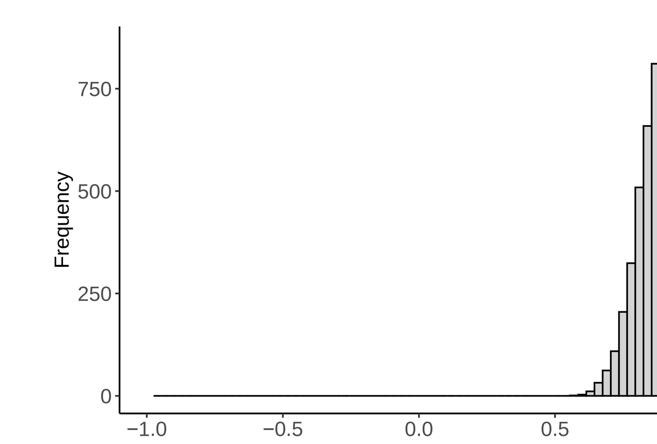


Figure 6: **No Identifiability problem** – Distribution of ARI when the values of π_{jk} are high.

How to detect identifiability problems? We propose to:

- bootstrap the data;
- fit an MLC-CUB model on each bootstrapped data set;
- compute the ARI between the original partition and the one obtained with bootstrapped data;
- look at the distribution of ARI.

Future works

In the future, we plan to:

- extend the model to a multilevel setting;
- use other models within the CUB framework;
- relax the independence assumption through copulas.

References

M. Corneli, C. Bouveyron, and P. Latouche. 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, 2020.

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