Italian Statistical Society Series on Advances in Statistics

Alessio Pollice · Paolo Mariani *Editors*

Methodological and Applied Statistics and Demography IV

SIS 2024, Short Papers, Contributed Sessions 2

SIS Società Italiana di Statistica

Springer

Italian Statistical Society Series on Advances in Statistics

This book series publishes volumes of peer-reviewed short papers presented at the scientific events such as conferences, seminars and workshops organized by the Italian Statistical Society (SIS) and its sections.

The Italian Statistical Society (Società Italiana di Statistica, SIS) was established in 1939 and ranks among the institutions of particular scientific relevance. SIS aims to promote scientific activities for the development of statistical sciences and carries out this task organizing scientific meetings and conferences, and by means of publications and partnerships at a national and international level.

Alessio Pollice · Paolo Mariani
Editors

Methodological and Applied Statistics and Demography IV

SIS 2024, Short Papers, Contributed Sessions 2

Springer

Editors
Alessio Pollice
Department of Economics and Finance
The University of Bari Aldo Moro
Bari, Italy

Paolo Mariani
Department of Economics, Management
and Statistics
University of Milano-Bicocca
Milan, Italy

ISSN 3059-2135 ISSN 3059-2143 (electronic)
Italian Statistical Society Series on Advances in Statistics
ISBN 978-3-031-64446-7 ISBN 978-3-031-64447-4 (eBook)
https://doi.org/10.1007/978-3-031-64447-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Clustering Multivariate Rating Data Within the CUB Framework

Matteo Ventura[1(✉)], Julien Jacques[2], and Paola Zuccolotto[1]

[1] Department of Economics and Management, University of Brescia, Brescia, Italy
{matteo.ventura,paola.zuccolotto}@unibs.it
[2] ERIC Laboratory, University Lyon 2, Lyon, France
julien.jacques@univ-lyon2.fr

Abstract. Among the models for the analysis of rating data, the CUB (Combination of discrete Uniform and shifted Binomial random variable) is particularly interesting because it gives an interpretation of the dual latent factors believed to influence the final decision of a rater: feeling and uncertainty. In essence, this model represents the distribution of final ratings as a combination of a shifted binomial and a uniform random variable. Within the framework provided by the CUB model, we propose a mixture of multivariate CUB models to cluster multivariate rating data, whose estimation is performed via the EM algorithm. To evaluate our approach, we conducted two simulations, showcasing the model's consistency in handling complex data structures and capturing underlying patterns. Our findings underscore the potential of this methodology in uncovering hidden structures within multivariate rating datasets, offering valuable insights in various research domains.

Keywords: Ordinal data · Model-based clustering · EM algorithm · Mixture model

1 The CUB Model

In marketing, psychological and social research, questionnaires serve as a common tool to evaluate latent traits such as the perceptions, opinions, and attitudes of respondents. These traits are often assessed using Likert-type rating scales, where the results are gathered as ordinal data, meaning that there is a natural ordering of the categories. Analyzing ordinal data can be challenging because it requires addressing the inherent properties of ordinal variables. To tackle this challenge, the CUB (Combination of discrete Uniform and shifted Binomial random variables) model has been introduced [2], and then extended by several researchers [6]. This model posits that the underlying decision-making process leading to respondents' final ratings is characterized by two latent components: the feeling and the uncertainty.

The feeling component represents the rational aspect, reflecting respondents' preferences for a particular item or attribute. On the other hand, the uncertainty component captures inherent indecision present in human choices.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 637–642, 2025.
https://doi.org/10.1007/978-3-031-64447-4_108

Given a scale with m categories, the rating $r = 1, \ldots, m$ is considered as the realisation of a mixture of a shifted Binomial and a discrete Uniform random variables, defined as follows:

$$P(r \mid \xi, \pi) = \pi P_{SB}(r \mid \xi) + (1 - \pi) P_U(m),$$

where $P_{SB}(r \mid \xi) = \binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}$, and $P_U(m) = \frac{1}{m}$. The former accounts for the feeling, measured by the so-called feeling parameter $1 - \xi \in [0, 1]$, and the latter accounts for the uncertainty, which is measured by the uncertainty parameter $1 - \pi \in (0, 1]$, that is, the mixing proportion of the mixture.

2 A Mixture of Multivariate CUB Models

Clustering is an important tool for researchers to discover hidden structures in data sets, and mixture models have been successfully applied to cluster data. However, ordinal data have received less attention compared to other types of data, especially in a model-based clustering context, where there are a few proposals. The two most notable models are a model-based clustering relying on a Binary Ordinal Search (BOS) algorithm [1] and clustMD, which assumes that ordinal data are a generalization of a latent Gaussian [5].

In this section, a mixture model for clustering multivariate rating data following the CUB paradigm is presented. The CUB model is univariate, however, clustering is usually performed with multivariate data. Therefore, we propose a clustering algorithm based on a mixture of multivariate CUB models, which are assumed to be conditionally independent.

Let $\mathbf{r} = (r_{ij})_{1 \leq i \leq n, 1 \leq j \leq J}$ be a multivariate ordinal variable, where the jth component is an ordinal variable with m_j categories; and let ω_k be the mixing proportion of cluster K, such that $\omega_k > 0$ and $\sum_{k=1}^{K} \omega_k = 1$.

Under the assumption of conditional independence of the variables, the marginal distribution of $\mathbf{r}$ is:

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

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

For mixture models there exists no close form for the ML estimators, therefore the parameters are estimated via the EM algorithm. To define the complete log-likelihood, two latent allocation variables are introduced in the model: $\mathbf{z} : (\mathbf{z}_k)_{1 \leq k \leq K}$ is a random variable distributed as a one-order Multinomial distribution, $\mathbf{z} \sim \mathcal{M}(1; \omega_1, \ldots, \omega_k)$, such that $z_{ik} = 1$ if the ith rater preferences come from cluster k, and $z_{ik} = 0$ otherwise. The second allocation variable $\mathbf{v} : (\mathbf{v}_j)_{1 \leq j \leq J}$, is dependent on the allocation variable z_{ik} and it is distributed as a Bernoulli with parameter π_{jk}. The variable $v_{ij} = 1$ 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:

$$\ell_c(\boldsymbol{\theta}; \mathbf{r}, \mathbf{z}, \mathbf{v}) = \sum_{1=1}^{n} \sum_{k=1}^{K} z_{ik} \Big\{ \ln(\omega_k) + \sum_{j=1}^{J} v_{ij} \big[\ln(\pi_{jk}) + \ln \big[P_{SB}(r_{ij} \mid \xi_{jk}) \big] \big] + \sum_{j=1}^{J} (1 - v_{ij}) \big[\ln(1 - \pi_{jk}) + \ln P_U(m_j) \big] \Big\}.$$

Starting from the initial set of parameters $\boldsymbol{\theta}^0 = (\boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\omega})^{(0)}$, the t-th iteration ($t \geq 0$) is the following:

– **E-step**: for all $1 \leq i \leq n$, $1 \leq j \leq J$, $1 \leq k \leq K$, the conditional probabilities of $\mathbf{z}$ are computed as follows:

$$\mathbb{E}(z_{ik} \mid \boldsymbol{r}_i; \theta^{(t)})$$

$$= \frac{\omega_k^{(t)} \prod_j \left[\pi_{jk}^{(t)} P_B(r_{ij} \mid \xi_{jk}^{(t)}) + (1 - \pi_{jk}^{(t)}) P_U(m_j) \right]}{\sum_{k'=1}^{K} \omega_{k'}^{(t)} \prod_j \left[\pi_{jk'}^{(t)} P_B(r_{ij} \mid \xi_{jk'}^{(t)}) + (1 - \pi_{jk'}^{(t)}) P_U(m_j) \right]}$$

$$= \tau_{ik}(\mathbf{r}_i; \theta^{(t)}) = \tau_{ik}^{(t)}.$$

Then, the conditional probabilities of the product $\mathbf{z} \cdot \mathbf{v}$ are computed:

$$\mathbb{E}(z_{ik} v_{ij} \mid r_{ij}; \theta^{(t)})$$

$$= \frac{\pi_{jk}^{(t)} P_B(r_{ij} \mid \xi_{jk}^{(t)})}{\pi_{jk}^{(t)} P_B(r_{ij} \mid \xi_{jk}^{(t)}) + (1 - \pi_{jk}^{(t)}) P_U(m_j)} \cdot \tau_{ik}(\mathbf{r}_i; \theta^{(t)})$$

$$= \nu_{ik}(r_{ij}; \theta^{(t)}) \cdot \tau_{ik}(\mathbf{r}_i; \theta^{(t)}) = \eta_{ijk}^{(t)}.$$

– **M-step**: update the estimation of the model parameters:

$$\pi_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n} \eta_{ijk}^{(t)}}{\sum_{i=1}^{n} \tau_{ik}^{(t)}},$$

$$\xi_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n} \eta_{ijk}^{(t)} (m_j - r_{ij})}{\sum_{i=1}^{n} \eta_{ijk}^{(t)} (m_j - 1)},$$

$$\omega_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)}}{n}.$$

The algorithm is stopped when a threshold is reached in the relative change of the log-likelihood.

3 Simulation Studies

In this section, a simulation study is performed to evaluate the performance of the EM algorithm. 100 datasets of size $n \in \{100, 1000\}$ have been simulated from a trivariate mixture ($J = 3$) of two CUB components ($K = 2$), each dimension having nine categories ($m_j = 9$). The set of parameters considered to generate the simulated data sets is shown in Table 1.

Table 1. Set of parameters chosen for generating the simulated data sets.

	$k = 1$			$k = 2$		
ω	0.40			0.60		
	$j = 1$	$j = 2$	$j = 3$	$j = 1$	$j = 2$	$j = 3$
π	0.60	0.60	0.50	0.70	0.80	0.70
ξ	0.30	0.20	0.10	0.70	0.80	0.70

3.1 Simulation Study 1

On each data set, the EM algorithm is run 20 times with uniform random starting parameters and setting the number of clusters equal to the real number ($K = 2$). The algorithm stops when the relative change in the log-likelihood is less than 10^{-4}. Finally, the run which provided the higher log-likelihood among the 20 has been considered as the final output of the EM algorithm. The results improve as the number of observations increases. The estimation of the parameter π_{jk} seems to be more difficult compared to the estimation of the other parameters. The boxplots of the parameter estimates dìstributions with a sample of size $n = 100$ and $n = 1000$ are reported in Figs. 1, 2 and 3.

3.2 Simulation Study 2

The second simulation aims to check if the model can detect the real number of clusters. As in the previous simulation, the EM algorithm is run 20 times on both the data sets with uniform random initialization, and it is stopped when the threshold on the relative change in the log-likelihood is reached. The EM algorithm is run by considering different numbers of clusters $K \in \{1, \ldots, 5\}$, with $K = 1$ meaning that there are no clusters and the model is simply a multivariate CUB. The estimated number of clusters corresponds to the one that leads to the lower BIC value. The performances in the detection of the real number of clusters are good with both sample sizes since the model detects two clusters in 97% of cases with the data set of size $n = 100$, and in 100% of cases in the data set of size $n = 1000$.

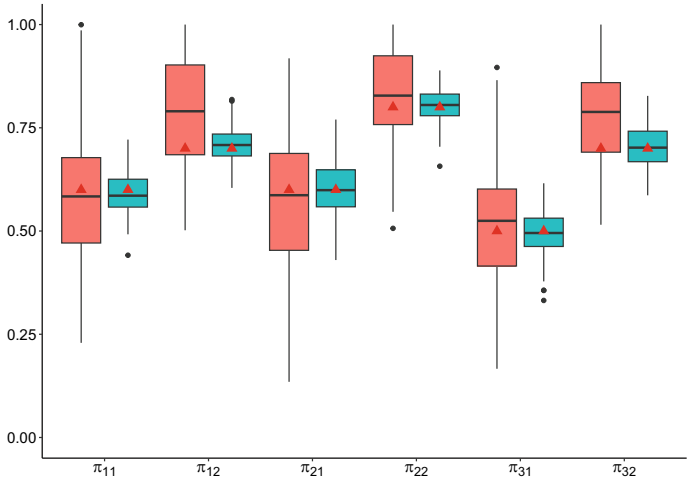

Fig. 1. Boxplots of the estimates of the parameters π_{jk} considering the sample with $n = 100$ observations (in red), and $n = 1000$ observations (in blue). The red triangles represent the true values of the parameters.

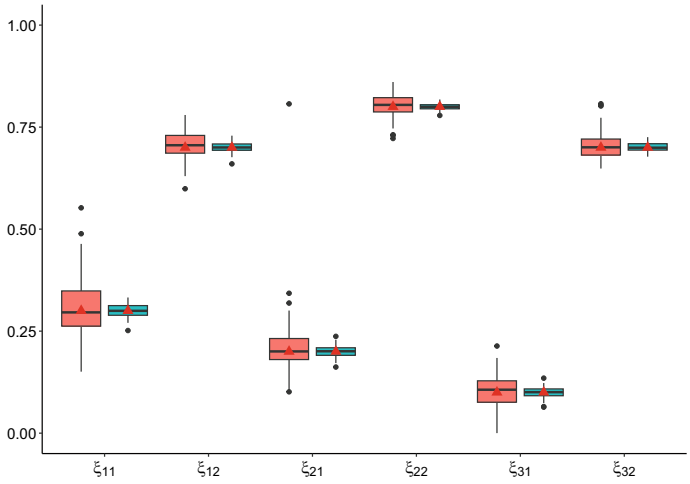

Fig. 2. Boxplots of the estimates of the parameters ξ_{jk} considering the sample with $n = 100$ observations (in red), and $n = 1000$ observations (in blue). The red triangles represent the true values of the parameters.

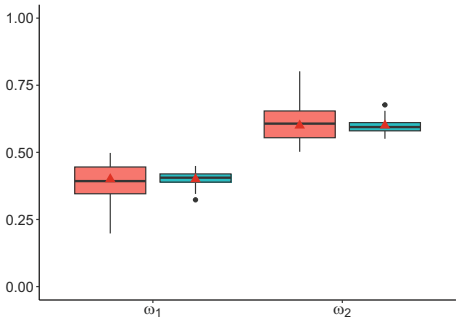

Fig. 3. Boxplots of the estimates of the parameters ω_k considering the sample with $n = 100$ observations (in red), and $n = 1000$ observations (in blue). The red triangles represent the true values of the parameters.

4 Conclusions

In this study, we introduced a mixture model for clustering rating data within the CUB framework. The simulation results demonstrate promising outcomes in estimating the mixing proportion and the feeling parameter. However, estimating the uncertainty parameter appears to be more challenging. Additional simulation scenarios will be explored to further assess the model's performance.

It is important to note that a mixture of CUB models is not identifiable due to the presence of the uniform component of different CUB models, which are not distinguishable [1,4]. However, we expect that the probability of a non-identifiable solution decreases as the number of variables in the data increases. Moreover, we intend to apply the model to real datasets to demonstrate its practical utility.

References

1. Biernacki, C., Jacques, J.: Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Stat. Comput. **26**, 929–943 (2016)
2. D'Elia, A., Piccolo, D.: A mixture model for preferences data analysis. Comput. Stat. Data Anal. **49**(3), 917–934 (2005)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B **39**(1), 1–38 (1977)
4. Grilli, L., Iannario, M., Piccolo, D., Rampichini, C.: Latent Class CUB models. Adv. Data Anal. Classif. **8**(1), 105–119 (2014)
5. McParland, D., Gormley, I.C.: Model based clustering for mixed data: clustMD. Adv. Data Anal. Classif. **10**(2), 155–169 (2016). https://doi.org/10.1007/s11634-016-0238-x
6. Agresti, A., Kateri, M.: The class of CUB models: statistical foundations, inferential issues and empirical evidence. Stat. Methods Appl. **28**(3), 445–449 (2019). https://doi.org/10.1007/s10260-019-00468-8