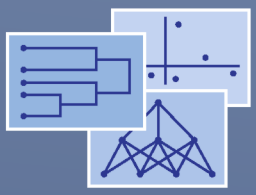
Studies in Classification, Data Analysis, and Knowledge Organization

Paula Brito · Francesco Denti · Francesca Greselin · Krzysztof Jajuga · Mariangela Zenga *Editors*

Navigating Complexity

Statistical Methods, Data Analysis, and Machine Learning for Actionable Insights



Springer

Studies in Classification, Data Analysis, and Knowledge Organization

Studies in Classification, Data Analysis, and Knowledge Organization is a book series which offers constant and up-to-date information on the most recent developments and methods in the fields of statistical data analysis, exploratory statistics, classification and clustering, handling of information and ordering of knowledge. It covers a broad scope of theoretical, methodological as well as application-oriented articles, surveys and discussions from an international authorship and includes fields like computational statistics, pattern recognition, biological taxonomy, DNA and genome analysis, marketing, finance and other areas in economics, databases and the internet. A major purpose is to show the intimate interplay between various, seemingly unrelated domains and to foster the cooperation between mathematicians, statisticians, computer scientists and practitioners by offering well-based and innovative solutions to urgent problems of practice.

Paula Brito · Francesco Denti ·
Francesca Greselin · Krzysztof Jajuga ·
Mariangela Zenga
Editors

Navigating Complexity

Statistical Methods, Data Analysis, and Machine Learning for Actionable Insights

Springer

Editors
Paula Brito
Faculty of Economics
University of Porto
Porto, Portugal

Francesca Greselin
Department of Statistics and Quantitative Methods
University of Milano-Bicocca
Milan, Italy

Mariangela Zenga
Department of Statistics and Quantitative Methods
University of Milano-Bicocca
Milan, Italy

Francesco Denti
Department of Statistics
University of Padova
Padua, Italy

Krzysztof Jajuga
Department of Financial Investments and Risk Management
Wrocław University of Economics and Business
Wrocław, Poland

ISSN 1431-8814 ISSN 2198-3321 (electronic)
Studies in Classification, Data Analysis, and Knowledge Organization
ISBN 978-3-032-32008-7 ISBN 978-3-032-32009-4 (eBook)
https://doi.org/10.1007/978-3-032-32009-4

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

Preface

The present volume, *Navigating Complexity - Statistical Methods, Data Analysis, and Machine Learning for Actionable Insights*, gathers the revised, refereed versions of short papers accepted for presentation at the 2026 Conference of the International Federation of Classification Societies (IFCS). The upcoming conference, themed "*Classification, Machine Learning and Statistical Methods: from Data to Discovery*", will take place in Milan, Italy from July 14 to July 16, 2026. The event is locally organized by a joint effort of University of Milano-Bicocca and University of Milan, and two hosting scientific societies: the Classification and Data Analysis Group (CLADAG) and the Italian Statistical Society (SIS).

Founded in 1985, the International Federation of Classification Societies (IFCS) is an international scientific organization dedicated to promoting mutual communication, cooperation, and the interchange of views among all those interested in the scientific principles, numerical methods, and practice of data science, data analysis, and classification. IFCS 2026 continues this proud tradition by serving as a premier platform for dissemination and collaboration. Indeed, this year's scientific program features over 260 accepted contributions, thoughtfully organized into thematic tracks, special sessions, contributed paper sessions, and poster presentations. We are excited to welcome over 300 scholars from countries of all continents to the upcoming event. We are particularly looking forward to the six keynote lectures to be addressed by Krzysztof Jajuga (Wrocław U of Economics and Business), Sugnet Lubbe (Stellenbosch U), Brendan Murphy (U College Dublin), Mohamed Nadif (U Paris Cité), Adrian Raftery (U Washington), and Laura Sangalli (Politecnico di Milano). Additionally, the conference program will include highly anticipated tutorials: "Machine learning with functional data: from statistical methods to deep architectures" by Alberto Suárez González (U Autónoma de Madrid) and "Hidden but powerful: practical insights into latent Markov models to discover dynamics in longitudinal data" by Fulvia Pennoni (U of Milano-Bicocca). A Data Challenge has also been organized to engage young students.

The Conference Scientific Program Committee is co-chaired by Francesca Greselin (U of Milano-Bicocca) and Silvia Salini (U of Milan), and includes representatives of the IFCS member societies: Krzysztof Jajuga (IFCS President, Wrocław U of Economics and Business), Rebecca Nugent (IFCS Past President, Carnegie Mellon U), Mark de Rooij (IFCS President Elect, Leiden U), Vladimir Batagelj (U of Ljubljana), Paula Brito (U do Porto), Sonya Coleman (Ulster U), Rosa Crujeiras (U de Santiago de Compostela), Jean Diatta (U of La Réunion), Luca Frigau (U of Cagliari), Raeesa Ganey (U of the Witwatersrand), Stefano Antonio Gattone (U di Chieti "G. d'Annunzio"), Aurea Grané (U Carlos III de Madrid), Julien Jacques (U Lumière Lyon 2), Joanna Landmesser-Rusek (Warsaw U of Life Sciences), Berthold Lausen (U of Essex), Agustin Mayo Iscar (U de Valladolid), Angelos Markos (Democritus U of Thrace), Raffaella Piccarreta (Bocconi U), Riccardo Rastelli (U College Dublin), Pieter Schoonees (Erasmus U Rotterdam), Luca Scrucca (U of Bologna), Piercesare Secchi (Politecnico di Milano), Gianluca Sottile

(U degli Studi di Palermo), Yoshikazu Terada (Osaka U), Cristina Tortora (San Jose State U), Tom Wilderjans (Leiden U), Adalbert Wilhelm (Constructor U). We wish to thank them, alongside the chairs of the thematic tracks, for their invaluable cooperation in evaluating submissions and shaping an outstanding scientific program prior to the event.

The local organization is chaired by Francesca Greselin (U of Milano-Bicocca). We cordially thank all members of the Local Organizing Committee – Luca Brusa (U of Milano-Bicocca), Andrea Cappozzo (Catholic U of Milan), Alessandro Casa (Free U of Bozen-Bolzano), Francesca De Battisti (U of Milan), Francesco Denti (U of Padova), Sónia Dias (Polytechnic Institute of Viana do Castelo), Andrea Marletta (U of Milano-Bicocca), Marta Nai Ruscone (U of Genova), Giorgia Zaccaria (U of Milano-Bicocca), Mariangela Zenga (U of Milano-Bicocca) – and everyone at University of Milan and University of Milano-Bicocca who has worked actively behind the scenes to prepare for this international gathering.

Reflecting this volume's title, *Navigating Complexity*, the selected papers collected here showcase modern methodologies and real-world applications designed to extract actionable insights from intricate data. The topics span a wide range of areas within statistics, machine learning, and data science. These include model-based clustering, Bayesian methods, anomaly detection, and predictive modeling. Novel and tailored models are proposed for complex data types, such as mixed-type data, compositional data, and functional data. Furthermore, advanced statistical concepts and cutting-edge artificial intelligence techniques are explored, underscoring the interdisciplinary effort required to address complex challenges in today's data-driven landscape.

Finally, the organizers and editors would like to express their deepest gratitude to all the authors for their commitment and outstanding contributions to this volume. We are especially grateful to the colleagues who served as reviewers, whose timely and dedicated work was decisive to the scientific quality of these proceedings. Their names are reported in the next page. We also thank our generous sponsors and Springer Nature for their support in producing this book ahead of the conference.

We look forward to a stimulating, fruitful, and highly engaging meeting at IFCS 2026.

June 2026

Paula Brito
Francesco Denti
Francesca Greselin
Krzysztof Jajuga
Mariangela Zenga

Acknowledgments

The editors are deeply grateful to the reviewers, whose contributions were essential to the scientific quality of these proceedings. They are listed below in alphabetical order:

Claudio Agostinelli
Leonardo Salvatore Alaimo
Alessandro Albano
Viviana Amati
Laura Anderlucci
Eleonora Arnone
Roberto Ascari
Luigi Augugliaro
Andrzej Bąk
Lynne Billard
Luca Brusa
Andrea Cappozzo
Alessandro Casa
Sonya Coleman
Dianne Cook
Pietro Coretto
Riccardo Corradin
Laura D'Angelo
Angela Maria D'Uggento
Lucio De Capitani
Grazyna Dehnel
Agnese Maria Di Brisco
Roberto Di Mari
Marco Di Marzio
Cinzia Di Nuzzo
José G. Dias
Pedro Duarte Silva
Andrzej Dudek
Peter Filzmoser
Alice Giampino
Sabrina Giordano
Erika Grammatica
Luca Greco
Patrick Groenen
Mia Hubert
Alina Jedrzejczak
Simona Korenjak-Cerne
Joanna Landmesser-Rusek
Berthold Lausen
Eufrasio Lima Neto
Rosaria Lombardo
Pawel Lula
Paolo Mariani
Andrea Marletta
Giovanna Menardi
Sara Milito
Krzysztof Najman
M. Rosário Oliveira
Andrea Ongarato
Francesco Palumbo
Andrea Panarotto
Leo Pasquazzi
Marcin Pełka
Alfonso Piscitelli
Marialuisa Restaino
Maurizio Romano
Gilbert Saporta
Pieter Schoonees
Piercesare Secchi
Andrzej Sokolowski
Andrea Sottosanti
Gero Szepannek
Paulo Teles
Salvatore Daniele Tomarchio
Cristina Tortora
Marek Walesiak
Adalbert Wilhelm

Contents

Mixture-Based Latent Trait Modeling for Biclustering Overdispersed Counts

Dalila Failli[1(✉)], Maria Francesca Marino[1], and Francesca Martella[2]

[1] Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Florence, Italy
dalila.failli@unifi.it

[2] Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy

Abstract. We propose a mixture of latent trait models for biclustering units and variables in the presence of multivariate, overdispersed count data. Units are grouped into homogeneous clusters called components through a finite mixture model. Simultaneously, within each component, variables are grouped into segments using a flexible specification of the linear predictor. Covariates are incorporated at the latent level of the model to account for their effect on component formation, while residual dependence among variables is captured by a multidimensional latent trait. A simulation study based on a varying number of units and variables is conducted to assess both clustering performance and the ability to correctly estimate model parameters.

Keywords: Co-clustering · Latent variable models · Negative Binomial · EM algorithm · Gauss-Hermite quadrature

1 Introduction

The joint partitioning of rows (i.e., units) and columns (i.e., variables) of a data matrix into homogeneous blocks is commonly known in the literature as biclustering, co-clustering, or block clustering (see, e.g., [5]). This approach is used to identify subsets of units that behave similarly across specific subsets of variables. Typical applications include genetics, where biclustering helps detect groups of genes showing similar expression patterns under certain experimental conditions, and market analysis, where the approach is used to segment customers according to their preferences for particular groups of products.

We propose a biclustering approach for multivariate, overdispersed count data based on a mixture of latent trait analyzers (MLTA; [9]) model. Units are partitioned into subsets (components) via a finite mixture of latent variable models, with component probabilities modeled as a function of observed covariates as in [7]. A multidimensional latent variable (trait) is also introduced to capture residual dependence among variables. Within each component, variables are partitioned into subsets (segments) via a flexible specification of the linear predictor, following an approach similar to that firstly introduced by [12], and

P. Brito et al. (Eds.): IFCS 2026, STUDIES CLASS, pp. 170–177, 2026.
https://doi.org/10.1007/978-3-032-32009-4_21

subsequently extended to the MLTA framework by [8] to deal with multivariate binary data.

In this work, we focus the attention on multivariate count responses. While the Poisson distribution is a standard choice for the modeling of count data, it often fails to accommodate the extra variability commonly observed in practice. To address this issue, we assume that, conditional on the latent trait and component membership, the responses of interest are independent of each other (local independence assumption) and follow a Negative Binomial (NB) distribution. Such a distribution introduces an additional dispersion parameter that allows the conditional variance to exceed the conditional mean, thus offering greater flexibility in the modeling of heterogeneous count processes.

A closely related approach is the Latent Block Model (LBM) for contingency tables proposed by [10], which provides a well-established framework for simultaneously clustering rows and columns by assuming block-wise homogeneity of the data matrix. While both approaches perform a joint clustering of units and variables, the proposed model introduces several important extensions compared to the LBM. First, as mentioned above, our proposal explicitly accommodates overdispersion in the data. Second, the model accounts for residual heterogeneity not fully captured by the block configuration. Third, the proposed model allows for cluster-specific partitions of the variables. In other words, for each row cluster, a potentially different grouping of variables is identified, enabling a more flexible representation of the data structure compared to the single global column partition assumed in the standard LBM framework. In addition, unlike the LBM, the proposed model allows for the inclusion of covariates to inform the clustering structure. Another related contribution is the model-based biclustering framework for overdispersed count data proposed in [2], which represents a special case of LBM that explicitly addresses overdispersion within a biclustering setting. Although this approach effectively relaxes the equi-dispersion assumption, it does not explicitly model residual heterogeneity beyond the bicluster structure. Moreover, it does not allow for cluster-specific partitions of variables or for clustering informed by covariates, thereby limiting its flexibility in more complex data scenarios.

The paper is organized as follows. In Sect. 2, we describe the model. Section 3 shows the results of the simulation study, while Sect. 4 contains concluding remarks and details further extensions of the approach.

2 The Proposed Model

Let Y_{ik} denote the k-th random variable and y_{ik} the corresponding observed value for the i-th unit, with $i = 1, \ldots, N$, and $k = 1, \ldots, R$. Each unit is assigned to one out of G unobserved clusters (components), identified by a discrete latent variable

$$\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})' \overset{iid}{\sim} \text{Multinomial}(1, (\eta_{i1}, \ldots, \eta_{iG})').$$

The generic element of this vector, z_{ig}, is equal to 1 if the i-th unit belongs to the g-th component (with probability η_{ig}), and is equal to 0 otherwise, for $g =$

$1, \ldots, G$, and $i = 1, \ldots, N$. Following the proposal by [7], we model component probabilities η_{ig} via the following multinomial logit specification

$$\eta_{ig} = \Pr(z_{ig} = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta}_g) = \frac{\exp\{\boldsymbol{x}_i' \boldsymbol{\beta}_g\}}{1 + \sum_{g'=2}^{G} \exp\{\boldsymbol{x}_i' \boldsymbol{\beta}_{g'}\}}, \quad g = 2, \ldots, G,$$

where $\boldsymbol{\beta}_g = (\beta_{0g}, \beta_{1g}, \ldots, \beta_{Jg})'$ is a $(J+1)$-dimensional vector of coefficients measuring the impact of covariates $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{iJ})'$ on η_{ig}.

For each unit $i = 1, \ldots, N,$, we also assume the existence of a D-variate, Gaussian latent variable, $\boldsymbol{u}_i$, with mean vector and covariance matrix equal to $\mathbf{0}$ and $\boldsymbol{I}_D$, respectively, to ensure identifiability.

The proposed model relies on a local independence assumption. Specifically, to deal with overdispersed counts, we assume that, conditional on $\boldsymbol{z}_i$ and $\boldsymbol{u}_i$, responses in the vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iR})'$ are independent of each other and follow a NB distribution. Therefore, the joint (conditional) density of the vector $\boldsymbol{y}_i$ (a realization of $\boldsymbol{Y}_i$) may be expressed as

$$\begin{aligned} f(\boldsymbol{y}_i \mid \boldsymbol{u}_i, z_{ig} = 1) &= \prod_{k=1}^{R} f(y_{ik} \mid \boldsymbol{u}_i, z_{ig} = 1) \\ &= \prod_{k=1}^{R} \binom{y_{ik} + \alpha_{gk} - 1}{\alpha_{gk} - 1} \left(\frac{\gamma_{gk}}{\gamma_{gk} + \alpha_{gk}}\right)^{y_{ik}} \left(\frac{\alpha_{gk}}{\gamma_{gk} + \alpha_{gk}}\right)^{\alpha_{gk}}, \end{aligned}$$

with location and overdispersion parameters denoted by γ_{gk} and α_{gk}, respectively.

To achieve a joint clustering of units and variables, we follow the proposal by [1,12], and [8] and assume that, within each component, the location parameter γ_{gk} is modeled according to the following specification:

$$\log\left[\gamma_{gk}(\boldsymbol{u}_i)\right] = b_g + \boldsymbol{a}_{gk}'(\boldsymbol{\zeta} + \boldsymbol{u}_i).$$

Here, b_g is a component-specific effect, while $\boldsymbol{\zeta}$, $\boldsymbol{u}_i$, and $\boldsymbol{a}_{gk}$ identify a set of D-dimensional vectors allowing for a partitioning of variables into one out of $D \leq R$ clusters (hereafter, segments). In particular, $\boldsymbol{a}_{gk} = (a_{gk1}, \ldots, a_{gkD})'$ is a D-dimensional row stochastic (parameter) vector, whose generic element, a_{gkd}, is equal to 1 if, within component g, the k-th variable belongs to segment d, and 0 otherwise. Furthermore, the parameter b_g represents the baseline (log-transformed) mean count for units in the g-th component, while $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_D)'$ represents a vector of fixed model parameters assumed to be constant across components. Therefore, the fixed effect ζ_d captures systematic deviations from the baseline values implied by b_g for those variables belonging to the d-th segment. Last, $u_{id} \in \boldsymbol{u}_i$ is a unit- and segment-specific latent trait that characterizes individual profiles by capturing residual heterogeneity in the transformed mean counts of variables belonging to segment d.

Denoting with $\boldsymbol{\Phi}$ the vector of all free model parameters and considering the modeling assumptions described above, the likelihood function for the proposed

model can be written as:

$$L(\boldsymbol{\Phi}) = \prod_{i=1}^{N} L_i(\boldsymbol{\Phi}) = \prod_{i=1}^{N} \left(\sum_{g=1}^{G} \eta_{ig} \int_{\mathbb{R}^D} f(\boldsymbol{y}_i \mid \boldsymbol{u}_i, z_{ig} = 1) f(\boldsymbol{u}_i) d\boldsymbol{u}_i \right). \quad (1)$$

Maximum likelihood estimation of $\boldsymbol{\Phi}$ can be performed via an Expectation-Maximization (EM) algorithm [6]. However, the multidimensional integral appearing in Eq. (1) does not admit a closed-form solution and approximation methods are thus needed. Here, we opt for a Gauss-Hermite quadrature approach [13]. Accordingly, the integral is approximated via a weighted summation over a pre-specified set of quadrature locations (say, Q), with given weights; see [8] for further details.

Note that within the Gauss-Hermite quadrature framework, the computational cost grows in the order of Q^D. However, in our framework, the latent dimension D directly corresponds to the number of segments and is therefore small, ensuring that the approximation remains computationally feasible and numerically stable. However, the exploration of alternative and more convenient approximation methods for higher-dimensional latent traits remains a direction for future work.

Regarding model selection, the proposal is estimated over a range of values for components G and segments D; the optimal specification is selected by minimizing a chosen information criterion, such as the Bayesian Information Criterion (BIC; [14]).

3 Simulation Study

To assess the performance of the proposed approach in terms of parameter estimation and clustering, we conduct a simulation study based on $S = 100$ replicates, each initialized with 100 random starting values to avoid the EM algorithm remains trapped in local maxima solutions, and different data sizes. Specifically, we simulate data from the proposed specification, by considering $G = 3$ components and $D = 2$ segments, along with a varying number of units ($N = 100, 500, 1000$) and variables ($R = 10, 20$). A single covariate drawn from a Gaussian distribution with unit mean and variance is included in the model for prior component probabilities – i.e., $x_{i1} \sim \mathcal{N}(1, 1)$. Model parameters are set as $\beta_{0g} = 0$, $\beta_{12} = -0.4$, $\beta_{13} = -0.9$, $\boldsymbol{b} = (-1.7, 0, 1.7)'$, and $\boldsymbol{\zeta} = (-2, 0.5)'$. These values are chosen to have moderate separation between clusters and segments, ensuring an identifiable block structure. For simplicity, the dispersion parameters are constrained to be constant across variables, so that they only depend on mixture components (i.e., $\alpha_{gk} = \alpha_g$, for all $k = 1, \ldots, R$) and these are set as follows: $\alpha_1 = 0.5, \alpha_2 = 1$, and $\alpha_3 = 1.5$. These values allows for varying overdispersion of counts across clusters.

3.1 Clustering Performance

The ability of the proposal to correctly partition units and variables is evaluated through the Adjusted Rand Index (ARI; [11]); this represents a standard measure

to evaluate the agreement between the true and the estimated clustering. Results of the simulation study are shown in Table 1 and demonstrate that, as the number of units and, above all, variables, increases, clustering performance improves.

Table 1. ARI mean and standard error (SE) across samples, for varying N and R.

		$R = 10$		$R = 20$	
		Unit	Variable	Unit	Variable
N	100	0.64 (0.11)	0.69 (0.25)	0.79 (0.09)	0.83 (0.24)
	500	0.66 (0.09)	0.78 (0.17)	0.83 (0.06)	0.97 (0.11)
	1000	0.69 (0.08)	0.79 (0.18)	0.84 (0.04)	0.99 (0.06)

3.2 Parameter Recovery

Table 2 reports the Relative Root Mean Squared Error (RRMSE) values for the model parameters $\boldsymbol{b}$, $\boldsymbol{\zeta}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ across different sample sizes (N) and numbers of variables (R). Values equal to zero indicate parameters that were perfectly estimated, while non-zero values reflect the typical relative error. Overall, the table shows that parameter estimates are generally accurate, with RRMSEs decreasing as either the number of units or the number of variables increases. For instance, the RRMSE for $\boldsymbol{\beta}$ decreases from $[0.48, 0.40]$ at $N = 100$ and $R = 10$ to $[0.09, 0.09]$ at $N = 1000$ and $R = 20$, demonstrating improved estimation precision with larger samples and a larger number of observed variables.

Table 2. RRMSE values across samples for $\boldsymbol{b}$, $\boldsymbol{\zeta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, for varying N and R.

		$R = 10$			
		$\boldsymbol{b}$	$\boldsymbol{\zeta}$	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$
N	100	$[0.38, 0, 0.22]$	$[0.13, 0]$	$[0.48, 0.40]$	$[0.47, 0, 0.35]$
	500	$[0.21, 0, 0.11]$	$[0.06, 0]$	$[0.33, 0.17]$	$[0.40, 0, 0.16]$
	1000	$[0.18, 0, 0.13]$	$[0.06, 0]$	$[0.27, 0.11]$	$[0.41, 0, 0.15]$
		$R = 20$			
		$\boldsymbol{b}$	$\boldsymbol{\zeta}$	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$
N	100	$[0.31, 0, 0.24]$	$[0.12, 0]$	$[0.37, 0.32]$	$[0.29, 0, 0.27]$
	500	$[0.12, 0, 0.15]$	$[0.07, 0]$	$[0.14, 0.14]$	$[0.12, 0, 0.22]$
	1000	$[0.08, 0, 0.08]$	$[0.05, 0]$	$[0.09, 0.09]$	$[0.01, 0, 0.02]$

3.3 A Comparison with Competitive Models

In this section, we compare the performance of the proposed model in terms of clustering accuracy with that of the LBM [10] and the overdispersed LBM (OD-LBM) [2]. Specifically, we considered two scenarios, based on $G = 3$ components, $D = 2$ segments, $N = 500$ units, and $R = 20$ variables. In the first scenario, data are generated according to the proposed model under the parameter configuration described in Sect. 3; in the second scenario, data are generated from an LBM specification. This study allows us to assess (i) the performance of the proposed model under correct specification and (ii) its behavior when the underlying structure originates from a different mechanism. In both scenarios, $S = 100$ datasets are generated and, for each method, clustering ability is evaluated via the ARI. The LBM is estimated using the `blockcluster` R package [4], while the OD-LBM is fitted using the `cobiclust` R package [3].

As shown in Table 3, under Scenario 1, both LBM and OD-LBM perform poorly, failing to capture the biclustering structure of the data. Note that, in the absence of overdispersion, the OD-LBM reduces to the LBM, which explains the nearly identical results observed between the two models. In Scenario 2, the LBM and OD-LBM achieve almost perfect ARI, reflecting that these models are correctly specified. On the other hand, despite the misspecification, the proposed model remains competitive, recovering most of the biclustering structure and thus indicating robustness of the results even in less ideal settings.

Table 3. ARI mean and standard error (SE) for the proposed model (PROP), the LBM, and the overdispersed LBM (OD-LBM) across the two simulation scenarios.

Scenario	Model	Unit ARI	Variable ARI
1	PROP*	0.83 (0.06)	0.97 (0.11)
	LBM	0.23 (0.05)	0.00 (0.02)
	OD-LBM	0.21 (0.05)	0.00 (0.02)
2	PROP	0.82 (0.21)	0.93 (0.25)
	LBM	0.97 (0.12)	1.00 (0.01)
	OD-LBM	0.99 (0.01)	1.00 (0.01)

* Results coincide with those in Table 1 for $N = 500$ and $R = 20$.

4 Conclusions

The proposed model allows for a joint partitioning of units and variables in the context of overdispersed counts, which arise quite often in real data applications, such as genomics, social sciences, and epidemiology. In detail, units are partitioned into components, and, in each of them, variables are partitioned

into segments, thanks to a very flexible finite mixture model specification. Furthermore, a continuous latent trait captures residual heterogeneity in individual profiles that is not explained by the joint partitioning of units and variables, also accounting for latent factors that may influence the observed counts. The model also accounts for the effect of covariates on the probability of belonging to the different components, thus enabling the integration of additional information that may influence the clustering structure. Furthermore, it is worth noticing that in the absence of meaningful row or column clusters, the model selection procedure may appropriately favor the simplest configuration, typically selecting $G = 1$ and/or $D = 1$, indicating no latent clustering structure.

The performance of the proposed model is demonstrated through a simulation study. Results of this study show that the model can be effectively used for dimensionality reduction, as well as for uncovering latent patterns in overdispersed count data. In particular, as the number of units and variables increases, the model is able to accurately recover the true parameter values, while maintaining good clustering performance for both units and variables. The simulation study further shows that the proposed method maintains strong performance even under model misspecification, highlighting its robustness and flexibility as an analytical tool of analysis.

While the current paper focuses on simulations, the proposed model can be readily applied to real-world overdispersed count data, such as genomic expression profiles, social network activity counts, or epidemiological incidence data. Future work could explore extensions to handle larger datasets, or adapt the framework to longitudinal or spatially structured count data.

References

1. Alfò, M., Marino, M.F., Martella, F.: Biclustering multivariate discrete longitudinal data. Stat. Comput. **34**, 42 (2024). https://doi.org/10.1007/s11222-023-10292-6
2. Aubert, J., Schbath, S., Robin, S.: Model-based biclustering for overdispersed count data with application in microbial ecology. Methods Ecol. Evol. **12**(6), 1050–1061 (2021). https://doi.org/10.1111/2041-210X.13582
3. Aubert, J.: cobiclust: Biclustering via Latent Block Model Adapted to Overdispersed Count Data. R package version 0.1.2 (2024). https://cran.r-project.org/web/packages/cobiclust/index.html
4. Bhatia, P. S., Iovleff, S., Govaert, G.: blockcluster: an R package for model-based co-clustering. J. Stat. Softw. **76**(9), 1–24 (2017). https://doi.org/10.18637/jss.v076.i09
5. Bock, H.: Simultaneous clustering of objects and variables. In: Tomassone, R. (ed.) Analyse des données et informatique, pp. 187–204. INRIA, Le Chesnay (1979)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. **39**(1), 1–38 (1977)
7. Failli, D., Marino, M.F., Martella, F.: Finite mixtures of latent trait analyzers with concomitant variables for bipartite networks: an analysis of COVID-19 data. Multivar. Behav. Res. **59**(4), 801–817 (2024). https://doi.org/10.1080/00273171.2024.2335391

8. Failli, D., Marino, M.F., Martella, F.: A novel approach for biclustering bipartite networks: an extension of finite mixtures of latent trait analyzers. J. Classif. **42**, 492–516 (2025). https://doi.org/10.1007/s00357-025-09500-x
9. Gollini, I., Murphy, T.B.: Mixture of latent trait analyzers for model-based clustering of categorical data. Stat. Comput. **24**, 569–588 (2014). https://doi.org/10.1007/s11222-013-9389-1
10. Govaert, G., Nadif, M.: Latent block model for contingency table. Commun. Stat. Theory Methods **39**(3), 416–425 (2010). https://doi.org/10.1080/03610920903140197
11. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)
12. Martella, F., Alfò, M.: A finite mixture approach to joint clustering of individuals and multivariate discrete outcomes. J. Stat. Comput. Simul. **87**(11), 2186–2206 (2017). https://doi.org/10.1080/00949655.2017.1322593
13. Pinheiro, J.C., Bates, D.M.: Approximations to the log-likelihood function in the nonlinear mixed-effects model. J. Comput. Graph. Stat. **4**(1), 12–35 (1995)
14. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)