Computational Statistics
https://doi.org/10.1007/s00180-023-01387-9

ORIGINAL PAPER

Row mixture-based clustering with covariates for ordinal responses

Kemmawadee Preedalikit[1] · Daniel Fernández[2,3] · Ivy Liu[4] · Louise McMillan[4] · Marta Nai Ruscone[5] · Roy Costilla[6]

Received: 8 August 2022 / Accepted: 3 July 2023
© The Author(s) 2023

Abstract

Existing methods can perform likelihood-based clustering on a multivariate data matrix of ordinal data, using finite mixtures to cluster the rows (observations) of the matrix. These models can incorporate the main effects of individual rows and columns, as well as cluster effects, to model the matrix of responses. However, many real-world applications also include available covariates, which provide insights into the main characteristics of the clusters and determine clustering structures based on both the individuals' similar patterns of responses and the effects of the covariates on the individuals' responses. In our research we have extended the mixture-based models to include covariates and test what effect this has on the resulting clustering structures. We focus on clustering the rows of the data matrix, using the proportional odds cumulative logit model for ordinal data. We fit the models using the Expectation-Maximization algorithm and assess performance using a simulation study. We also illustrate an application of the models to the well-known arthritis clinical trial data set.

Keywords Cluster analysis · EM algorithm · Mixture models · Ordinal responses · Proportional odds model

✉ Daniel Fernández
daniel.fernandez.martinez@upc.edu

[1] School of Science, University of Phayao, Phayao 56000, Thailand

[2] Serra Húnter fellow. Department of Statistics and Operations Research (DEIO), Universitat Politècnica de Catalunya · BarcelonaTech (UPC), 08028 Barcelona, Spain

[3] Institute of Mathematics of UPC - BarcelonaTech (IMTech), 08028 Barcelona, Spain

[4] School of Mathematics and Statistics, Victoria University of Wellington, Wellington 6140, New Zealand

[5] DIEC - Department of Economics, University of Genoa, 16126 Genoa, Italy

[6] AgResearch Limited, Ruakura Research Centre, Hamilton 3214, New Zealand

Published online: 22 July 2023

1 Introduction

A well-known definition of an ordinal variable says it is one characterized by a categorical data scale, which describes an order showing differing degrees of dissimilarity (Agresti 2014). Thus, although ordinal variables are affected by the distances among their ordinal categories, those distances are not known. For example, in a questionnaire, the answers based on a Likert scale could be labelled as strongly disagree, disagree, neutral, agree, and strongly agree, and these are frequently coded as an equally-distanced 1–5 scale, but they could be coded using any other increasing sequence of numerical values. Ordinal scales are very common in a wide range of areas such as medical studies, ecology, and marketing.

Cluster analysis is the study of techniques to classify a set of related objects into the same cluster (Everitt et al. 2011) and can be applied to identify groups, patterns, or clusters in a data set. Clustering is used for a wide range of applications, in fields including business, biology, psychology, and medicine. A couple of recent examples of this are an application to gene microarray data proposed by Rocci and Vichi (2008) and an application in the analysis of genomic abnormality data, in which the developmental patterns of different types of tumours are used to identify clusters of tumours (Hoff 2005).

Many different approaches to clustering have been developed. The earliest approaches use partition optimization; the most common method is the k-means clustering (MacQueen 1967; Hartigan and Wong 1979). Several authors have proposed extensions to this approach [see e.g. Vichi (2001) and Rocci and Vichi (2008)]. Moreover, the objects can be clustered in a hierarchical way, gradually agglomerating objects into larger and larger clusters (Ward 1963; Johnson 1967). All approaches listed above are based on mathematical distance metrics and therefore statistical inferences, model selection procedures, and goodness-of-fit assessments cannot be easily applied due to the lack of an underlying probability model (Everitt et al. 2011; Fernández et al. 2016).

Cluster analysis based on finite mixture models (Peel and McLachlan 2000) assumes that variables in the data matrix arise from mixtures of statistical distributions, with each cluster corresponding to one component of the mixture. The estimated parameters for those distributions are those that have the maximum likelihood based on the observed data. Likelihood-based methods include those proposed by McLachlan and Basford (1988), Peel and McLachlan (2000), Böhning et al. (2007), and Melnykov and Maitra (2010), among others. More recently, Govaert and Nadif (2010) and Pledger and Arnold (2014) proposed an approach via finite mixtures for binary and count data using Bernoulli or Poisson building blocks. Other authors have introduced clustering algorithms specifically for ordinal data: see e.g. Giordan and Diana (2011), Biernacki and Jacques (2016), Ranalli and Rocci (2016), Matechou et al. (2016), and Fernández et al. (2016, 2019). Matechou et al. (2016) proposed a mixture-based biclustering solution relying on the proportional odds assumption of the cumulative logit model (McCullagh 1980). Fernández et al. (2016) developed an equivalent model-based clustering approach using the ordered stereotype model (Anderson 1984), although this approach assumes that there are no covariates available. Furthermore, methods that cluster observations using both ordinal and continuous variables simultaneously, such as the approach proposed by Ranalli and Rocci (2017), should also be mentioned and compared in the context of our proposed method.

Unlike distance-based methods, which only determine which objects should be clustered together, likelihood-based methods can additionally describe the properties of each cluster, based on the estimated parameters, and can also estimate the probability of each object being allocated to each cluster. Additionally, the mixture-based approaches for ordinal responses introduced above are focused on finding cluster structures based only on the matrix of ordinal responses, and assume that no associated covariates are available. Any available covariates can be analyzed alongside the clustering results, to assist with interpretation of the cluster structures, even though there has been no reference to the covariates during the clustering process, but actually incorporating covariates in the clustering process could lead to different estimated clustering structures, and a different estimate of the number of clusters. Generally speaking, if a model with covariates is estimated, subjects tend to be clustered according to their responses and covariate effects. Therefore, it is desirable to make available covariates endogenous to the clustering process to improve interpretation of the main characteristics of the clusters (Murphy and Murphy 2020).

Other studies have investigated the associations between clustering structures and covariates. Gudicha and Vermunt (2013) described several methods for clustering categorical responses via a three-step approach: (1) estimate the mixture model; (2) assign subjects to clusters; (3) regress cluster assignments on the covariates. Another proposal, the cluster-weighted model (CWM) approach, fits the joint distribution of a random vector composed of a response variable and a set of covariates (Ingrassia et al. 2012; Lamont et al. 2016). Ingrassia et al. (2015) also introduced a version of the CWM for mixed-type covariates that assumes continuous covariates arise from Gaussian distributions. Finally, several methods in the literature use the mixture of experts (MoE) paradigm in which the parameters of the mixture are modelled as functions of fixed, potentially mixed-type, covariates (Formann 1992; Jacobs et al. 1991; Murphy and Murphy 2020).

Our approach to mixture-based clustering involves constructing an additive linear model of parameters, connected to the response data via a link function. Additional terms such as covariates may easily be added to the linear predictor. To the best of our knowledge, Fernández et al. (2019) introduced this formulation of model-based clustering for ordinal data with covariates, but the performance of these covariate methods and, more importantly, their influence on the resulting clustering structures, have not been documented so far. The main purpose of this article is to extend such models to include covariates and allow them to affect the detection of cluster structures. Moreover, we are also interested in comparing how the resulting clustering structures compare to those obtained without covariates, and how these changes may affect the interpretation of the results.

We will focus on extending the one-dimensional clustering approach proposed in Matechou et al. (2016). This approach models ordinal response data using the proportional odds assumption of the cumulative logit model (hereafter referred to as the "proportional odds model"). We will include covariates directly in the linear predictor.

Our approach to clustering follows the constructivist approach described by Hennig (2015), but with an interest in realist clustering: we think there are many scenarios where patterns in the data can be simplified by identifying clusters of observations that follow similar patterns, but if there is a real structure in the data, then we wish to determine that structure. There are many real-world scenarios that we can model as a response variable being affected by predictor variables, and in some of those scenarios, certain groups of observations may have different patterns of response to the predictors than other groups of observations. If those groups have already been identified, then we might attempt mixed model analysis or multilevel modelling; but if the groups have not already been identified, then the method we propose here provides a pathway to detecting these groupings of response patterns. So our approach could be seen as a bridge between regression modelling and cluster analysis.

The rest of the article is organised as follows. Section 2 introduces the one-dimensional clustering models and their formulation. Section 3 describes the measures used to compare different clustering structures. Section 4 uses a simulation study to assess the performance of the method, and Sect. 5 applies the method to a real-world application: the well-known *arthritis clinical trial* data. Section 6 describes our conclusions.

2 The row clustering model

When the data are in matrix form, clustering of rows is called row clustering. In this section, we present the row clustering formulation for finite mixtures based on the proportional odds model. This closely follows the model formulations in Matechou et al. (2016) and Fernández et al. (2019). We decided to focus on row clustering because it is more common to have covariates linked to observations (rows) than to variables (columns).

2.1 Model formulation

We consider a set of n subjects and m ordinal response variables, each with q possible ordinal response categories. Thus, data can be represented by an $n \times m$ matrix $\mathbf{Y}$ with ordinal entries y_{ij}. The row cluster index r ($r = 1, \dots, R$) represents the number of the row cluster and the symbol $i \in r$ indicates that row i is allocated to row cluster r. We shall assume that all rows belonging to the same row cluster r have ordinal responses driven by the same row cluster effect, i.e. that there are no individual row effects. In the case of the proportional odds model where the effect of rows on the response is considered, the probability that y_{ij} takes category k, when row i is in row cluster r, is defined by

$$P[y_{ij} = k | i \in r] = \theta_{ijrk},$$

where $i = 1, \dots, n$, $j = 1, \dots, m$ and $k = 1, \dots, q$ with $\sum_{k=1}^{q} \theta_{ijrk} = 1$ for a given i, j and r. This can be expressed using linear predictor terms as

$$\operatorname{logit}\left(\sum_{h=1}^{k}\theta_{ijrh}\right)=\eta_{ijrk}=\mu_k-(\alpha_r+\beta_j+\gamma_{rj}), \tag{1}$$

where the parameters $\{\mu_k\}$ are the cutpoints. $\{\alpha_r\}$ and $\{\beta_j\}$ indicate the effects of row cluster r and column j, respectively, and $\{\gamma_{rj}\}$ represent the associations between different row clusters and individual columns. Corner-point or sum-to-zero constraints on $\{\alpha_r\}$, $\{\beta_j\}$ and $\{\gamma_{rj}\}$ must be included to avoid identifiability problems and the monotonically increasing constraint $\mu_1 < \mu_2 < \ldots < \mu_q (= \infty)$ is included to capture the ordinal nature of the responses. The (unknown) proportion of rows in each row cluster r is defined as $\{\pi_1, \ldots, \pi_R\}$, with $\sum_{r=1}^{R} \pi_r = 1$.

In a simpler model with clustering of rows, the clustering is solely based on the patterns of responses of the rows (observations/subjects) without considering the information present in the covariates. For instance, let's consider a hypothetical example of a matrix of subjects answering a set of 5-level Likert-scale questions from a self-report questionnaire, which intends to measure the degree of suffering in patients diagnosed with cancer. If the covariate information is not incorporated in the clustering process, resulting clusters would only be based on the patterns of responses of the patients. For example, the clusters may be categorized as low scores, middle scores, and high scores, based solely on the responses. However, when the covariate information, such as type of cancer, treatment dose, initial tumor burden, size of the tumor, gender, and age, is included in the clustering process, the resulting clusters may differ. This is because patients with equal or similar values in the covariates are assumed to be *a priori* more likely to co-cluster than others. For instance, patients with larger tumor sizes may tend to be clustered together, regardless of their responses to the questionnaire. This motivational example is based on the one given in Müller et al. (2011).

We now define the model formulation of row clustering using the proportional odds model, with additional covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, as follows,

$$\operatorname{logit}\left(\sum_{h=1}^{k}\theta_{ijrh}\right)=\eta_{ijrk}=\mu_k-(\alpha_r+\beta_j+\gamma_{rj}+\mathbf{x}_i^T\boldsymbol{\delta}_r), \tag{2}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are a set of p covariates associated with row i of the data matrix; these covariates can be categorical or continuous. The parameters $\{\delta_r\}$ represent the effects of the covariates; we assume these effects are the same for all rows in the same row cluster r. When fitting this model, the subjects will be clustered according to both their response patterns and the values of their covariates, which may lead to different estimates of cluster assignment.

Considering the simplest row clustering model, without column effects, the proportional odds model without covariates can be expressed as

$$\operatorname{logit}\left(\sum_{h=1}^{k}\theta_{ijrh}\right)=\eta_{ijrk}=\mu_k-\alpha_r, \tag{3}$$

where the number of parameters, including the $R-1$ independent values of π_r, is $q+2R-3$.

Adding p covariates into model (3), we obtain

$$\text{logit}\left(\sum_{h=1}^{k}\theta_{ijrh}\right) = \eta_{ijrk} = \mu_k - (\alpha_r + \mathbf{x}_i^T\boldsymbol{\delta}_r), \tag{4}$$

where there are now $q + (p + 2)R - 3$ parameters in the model.

The row clustering model with individual column effects can be expressed as

$$\text{logit}\left(\sum_{h=1}^{k}\theta_{ijrh}\right) = \eta_{ijrk} = \mu_k - (\alpha_r + \beta_j), \tag{5}$$

where the number of parameters, including π_r, is $q + 2R + m - 4$.

Adding p covariates into model (5), we obtain the following model

$$\text{logit}\left(\sum_{h=1}^{k}\theta_{ijrh}\right) = \eta_{ijrk} = \mu_k - (\alpha_r + \beta_j + \mathbf{x}_i^T\boldsymbol{\delta}_r), \tag{6}$$

where the number of parameters, including π_r, is $q + (p + 2)R + m - 4$.

Models (3) and (4) will be used in the simulation and application section to compare the clustering structure.

2.2 Estimation of the parameters

The Expectation-Maximization (EM) algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997) is a well-known iterative procedure to compute maximum likelihood estimates in the presence of missing or incomplete data. In the suite of models introduced in the previous section, the actual row cluster memberships, i.e. the allocation of rows into row clusters, are unknown or missing. Thus, the EM algorithm is a natural approach to fit these models. Previous examples of this approach include Bernoulli and Poisson distributions (Pledger and Arnold 2014), the proportional odds model (Matechou et al. 2016), and the ordered stereotype model (Fernández et al. 2016).

We have modified the EM algorithm used in Matechou et al. (2016) and Fernández et al. (2016) to incorporate covariates. Assuming the local independence assumption, where variables within a row are conditionally independent of each other given the row's cluster membership (Clogg 1988), the incomplete data likelihood function can be expressed as

$$L\big(\Theta|\{y_{ij}\},\{\mathbf{x}_i\}\big) = \prod_{i=1}^{n}\left[\sum_{r=1}^{R}\pi_r\prod_{j=1}^{m}\prod_{k=1}^{q}\big(\theta_{ijrk}|\{\mathbf{x}_i\}\big)^{I(y_{ij}=k)}\right], \tag{7}$$

which sums over all possible partitions of rows into R clusters. $\mathbf{Y} = \{y_{ij}\}$ is the data matrix corresponding to the observed responses, and $\mathbf{X} = (\mathbf{x}_1^T, \dots, \mathbf{x}_n^T)^T$ is the matrix of p covariates for all n rows Θ contains all unknown parameters and π_r is the *a priori* row membership probability of row i.

The incomplete data likelihood is difficult to optimize numerically because it does not have a simple form. Therefore, it is more natural to work with the complete data likelihood which we define below.

Let $\mathbf{Z} = \{Z_{ir}\}$ be a set of random vectors corresponding to the missing information, i.e., the unknown row cluster memberships. $Z_{ir} = 1$ if row i is in row cluster r and 0 otherwise. Thus, $\sum_{r=1}^{R} Z_{ir} = 1$ for all i. We can then suppose a complete data set exists, $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$ and the complete data log-likelihood function can be defined as

$$\ell_c(\Theta|\{y_{ij}\}, \{\mathbf{x}_i\}, \{Z_{ir}\}) = \sum_{i=1}^{n} \sum_{r=1}^{R} Z_{ir} \log(\pi_r) + \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{r=1}^{R} \sum_{k=1}^{q} Z_{ir} I(y_{ij} = k) \log\left(\theta_{ijrk}|\{\mathbf{x}_i\}\right) \tag{8}$$

Using the previous equation, we can now determine the E-step and M-steps of the EM algorithm. Given the latest estimates $\widehat{\Theta}^{(t-1)}$ from the previous iteration, the expected value of the complete data log-likelihood over Z_{ir}, given the observed data $\{\mathbf{x}_i\}$ and $\{y_{ij}\}$, becomes

$$\begin{aligned} Q(\Theta|\widehat{\Theta}^{(t-1)}) =& E_{\{Z_{ir}\}|\{y_{ij}\},\{\mathbf{x}_i\},\widehat{\Theta}^{(t-1)}}[\ell_c(\Theta|\{y_{ij}\}, \{\mathbf{x}_i\}, \{Z_{ir}\})] \\ =& \sum_{i=1}^{n} \sum_{r=1}^{R} \log(\pi_r^{(t-1)}) E[Z_{ir}|\{y_{ij}\}, \{\mathbf{x}_i\}, \widehat{\Theta}^{(t-1)}] \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{r=1}^{R} \sum_{k=1}^{q} I(y_{ij} = k) \log\left(\widehat{\theta}_{ijrk}^{(t-1)}|\{\mathbf{x}_i\}\right) E[Z_{ir}|\{y_{ij}\}, \{\mathbf{x}_i\}, \widehat{\Theta}^{(t-1)}]. \end{aligned} \tag{9}$$

In the E-step, we use the latest parameter estimates Θ to find the expected values of Z_{ir}. The expected value of Z_{ir}, a Bernoulli variable, is the posterior probability of individual i being in cluster r given the observed data. Therefore, using Bayes' rule, we can compute it as

$$\begin{aligned} \hat{Z}_{ir}^{(t)} &= P[Z_{ir} = 1|\{y_{ij}\}, \{\mathbf{x}_i\}, \widehat{\Theta}^{(t-1)}] \\ &= \frac{P(\{y_{ij}\}|Z_{ir} = 1, \widehat{\Theta}^{(t-1)}, \{\mathbf{x}_i\}) P(Z_{ir} = 1)}{\sum_{\ell=1}^{R} P(\{y_{ij}\}|Z_{i\ell} = 1, \widehat{\Theta}^{(t-1)}, \{\mathbf{x}_i\}) P(Z_{i\ell} = 1)} \\ &= \frac{\hat{\pi}_r^{(t-1)} \prod_{j=1}^{m} \prod_{k=1}^{q} (\hat{\theta}_{ijrk}^{(t-1)}|\{\mathbf{x}_i\})^{I(y_{ij}=k)}}{\sum_{\ell=1}^{R} \{\hat{\pi}_\ell^{(t-1)} \prod_{j=1}^{m} \prod_{k=1}^{q} (\hat{\theta}_{ij\ell k}^{(t-1)}|\{\mathbf{x}_i\})^{I(y_{ij}=k)}\}}. \end{aligned} \tag{10}$$

Then, we substitute this expected value of Z_{ir} in the complete data log-likelihood (9) at iteration t to complete the E-step,

$$\hat{Q}(\Theta|\Theta^{(t-1)}) = \sum_{i=1}^{n} \sum_{r=1}^{R} \hat{Z}_{ir}^{(t)} \log(\pi_r^{(t-1)}) + \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{r=1}^{R} \sum_{k=1}^{q} \hat{Z}_{ir}^{(t)} I(y_{ij} = k) \log\left(\hat{\theta}_{ijrk}^{(t-1)}|\{\mathbf{x}_i\}\right). \tag{11}$$

At the M-step, we maximize equation (11) obtained in the E-step with respect to π_r and Θ. The M-step estimates for finite mixture models can be calculated in two parts: the row-cluster proportions $\hat{\pi}_1, \dots, \hat{\pi}_R$ and the parameters $\widehat{\Theta}$. To find the estimates of π_r, following Fernández et al. (2016), we replace the conditional expectation (10) in the following expression for the iteration t,

$$\hat{\pi}_r^{(t)} = \frac{1}{n} \sum_{i=1}^{n} E[Z_{ir} | \{y_{ij}\}, \{\mathbf{x}_i\}, \Theta^{(t-1)}] = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ir}^{(t)}. \tag{12}$$

Similarly, to find the estimate of parameters Θ in the second part of (11), the derivative of the second term can be taken with respect to Θ. However, this has no simple analytical solution; we need to find the conditional expectation of the complete data log-likelihood of equation (9) using numerical maximization.

We then iterate the E-step and the M-step until we reach convergence. There are various convergence conditions that can be specified; we will use the convergence criterion based on the incomplete likelihood: we will iterate until the absolute difference between the incomplete log-likelihoods at two consecutive iterations, relative to the likelihood at the latest iteration, is close to zero. That is,

$$\frac{||L(\Theta^{(t+1)} | \{y_{ij}\}) - L(\Theta^{(t)} | \{y_{ij}\})||}{||L(\Theta^{(t)} | \{y_{ij}\})||} \approx 0. \tag{13}$$

At the end of the process, we have estimates for the posterior probabilities of cluster membership for each row, and these may be between 0 and 1. We will assume each observation is assigned to the group having the highest posterior probability.

We implemented the EM algorithm described above for the proportional odds model including clustering via finite mixtures and set up the simulation study by using the statistical software **R** 4.0.2 (**R** Development Core Team, 2019). The numerical maximization part of the M-step was carried out using the quasi-Newton method L-BFGS-B provided as an option in the predefined **R** function `optim()`. We used the default settings for all other control parameters. Alternative functions for maximum likelihood estimation of the cumulative version of the proportional odds model, assuming δ_{rj}, could be explored and used to intend to simplify the implementation process.

We remark that an inherent drawback in mixture modelling is that the associated likelihood surface may be multimodal. We therefore tried different starting points, covering a comprehensive range of parameter values, to avoid being locked into a local maximum. We reran the EM algorithm 10 times with random starting points and kept the run with the highest log-likelihood. In preliminary tests, we ran experiments testing up to 100 random starting points and found that 10 starting points were sufficient to avoid convergence to local optima. Finally, to ensure that this approach does not affect any final estimates, we used the resulting maximum likelihood estimation of the complete data likelihood using the EM algorithm as starting points (Fernández et al. 2016) to numerically maximize the incomplete data log-likelihood (7).

3 Measures to compare clustering structures

This section discusses three popular measures for comparing clustering structures: the Adjusted Rand Index (ARI 1985), the variation of information (VI 2005), and the normalised information distance (NID 2005). Comparing clustering structures can be challenging due to the “label-switching problem" where different labels can result in identical clusters. To address this issue, the measures used in this section do not rely on cluster labels, but instead consider pairs of rows that are clustered together. The Rand Index (RI 1971) measures similarity between clustering structures based on how data points are assigned to clusters, but it can have limitations in comparing replicability of different classifications. The ARI is an adjustment of the Rand index that corrects for chance with respect to the null hypothesis and ranges from 0 (totally independent structures) to 1 (identical structures). The VI measures the distance between partitions of the same dataset using concepts of entropy and information (Meila 2007), and the normalised VI (NVI 2005) is used to bound it between 0 and 1 for comparability with the ARI. The NID is another information criterion bounded between 0 and 1, and both the NVI and NID have values of 0 indicating identical clustering structures and values of 1 indicating totally independent structures. To simplify interpretation, the unit complements of the NVI and NID (1-NVI and 1-NID) are used in this section.

4 Simulation study

We set up a small scale simulation study to test, in a diverse range of scenarios, how reliably we were able to estimate both the parameters of our proposed row clustering model (4) and the cluster allocations, using the EM algorithm. We are not testing model selection here: we simulate data sets and then fit the correct model to those data. This study is closely related to the one in Fernández et al. (2016).

We simulated the simplest covariate model (4), with only a single covariate x_i and no column effects. We designed two possible main scenarios for the true model by varying the values of covariate effect parameters for the different clusters, $\{\delta_r\}$. Scenario 1 is designed with both negative and positive covariate effects, which means that different clusters could have dramatically different covariate effects. Scenario 2, by contrast, has only positive covariate effects, which is likely to make it more difficult to fit the cluster parameters, because the different clusters are more likely to produce similar response data than they were in Scenario 1.

The simulation program was written in **R**, and we did not observe any issues with its convergence. For each scenario, we ran several cases varying the following features:

- Sample size: $n = 100, 1000$
- Number of response categories: $q = 3, 4, 5, 6$
- Number of columns: $m = 3, 5, 10$
- Number of row clusters: $R = 3, 5$

- Distribution of covariates: Normal ($N(0, 1),$), Binomial ($Bin(1, 0.5)$)

In total, we ran 96 cases within each scenario. We generated 2000 replicate datasets for each combination of features using model (4) and calculated maximum likelihood estimates (MLEs) of the model parameters and their standard errors for each replicate. We then compared the estimated parameter values with the true parameter values and assessed the agreement between the true and estimated clustering structures using indices such as Adjusted Rand Index (ARI), 1-Normalized Variation of Information (1-NVI), and 1-Normalized Information Distance (1-NID). To report the results, we computed the mean of both the estimated model parameters and their corresponding standard errors using the 2000 simulated datasets.

4.1 Scenario 1

We simplified the study by using equal proportions of rows in each cluster: $(\pi_1, \dots, \pi_R) = (1/R, \dots, 1/R)$. The cutpoint values $\{\mu_k\}$ were chosen from a quantile function for the logistic distribution. Therefore, the cutpoint values are $\{\mu_1 = \log(1/2), \mu_2 = \log(2)\}$ when $q = 3$, and $\{\mu_1 = \log(1/4), \mu_2 = \log(2/3), \mu_3 = \log(3/2), \mu_4 = \log(4)\}$ when $q = 5$. We used evenly distributed values for the row cluster effect parameters α_r, with the corner-point constraint that $\alpha_1 = 0$.

Table 1 summarizes the average absolute bias and their corresponding standard errors for each parameter over 2000 simulations when the fitted models are model (3) and (4). In all cases, the estimated parameters of model (4) are close to the true values due to a small bias, and as expected, the variability decreases with increasing sample size n. We also remark that the value of the standard error decreases as the number of ordinal categories increases (see additional results in Tables 4 and 5, Figs. 4 and 5, Appendix A). We believe this might be due to the fact that as the number of ordinal categories increases, the response data becomes more continuous and the responses contain more information. On the other hand, the estimated parameters of model (3) perform poorly, i.e., they are very far from the true parameter values despite having modest standard errors (see additional results in Tables 6 and 7, Figs. 6 and 7, Appendix A).

4.2 Scenario 2

The setting of Scenario 2 is configured in the same way as the one from Scenario 1, apart from the specific values of the covariate effect parameters $\{\delta_r\}$. Scenario 2 has covariate effects in the same direction for all the clusters.

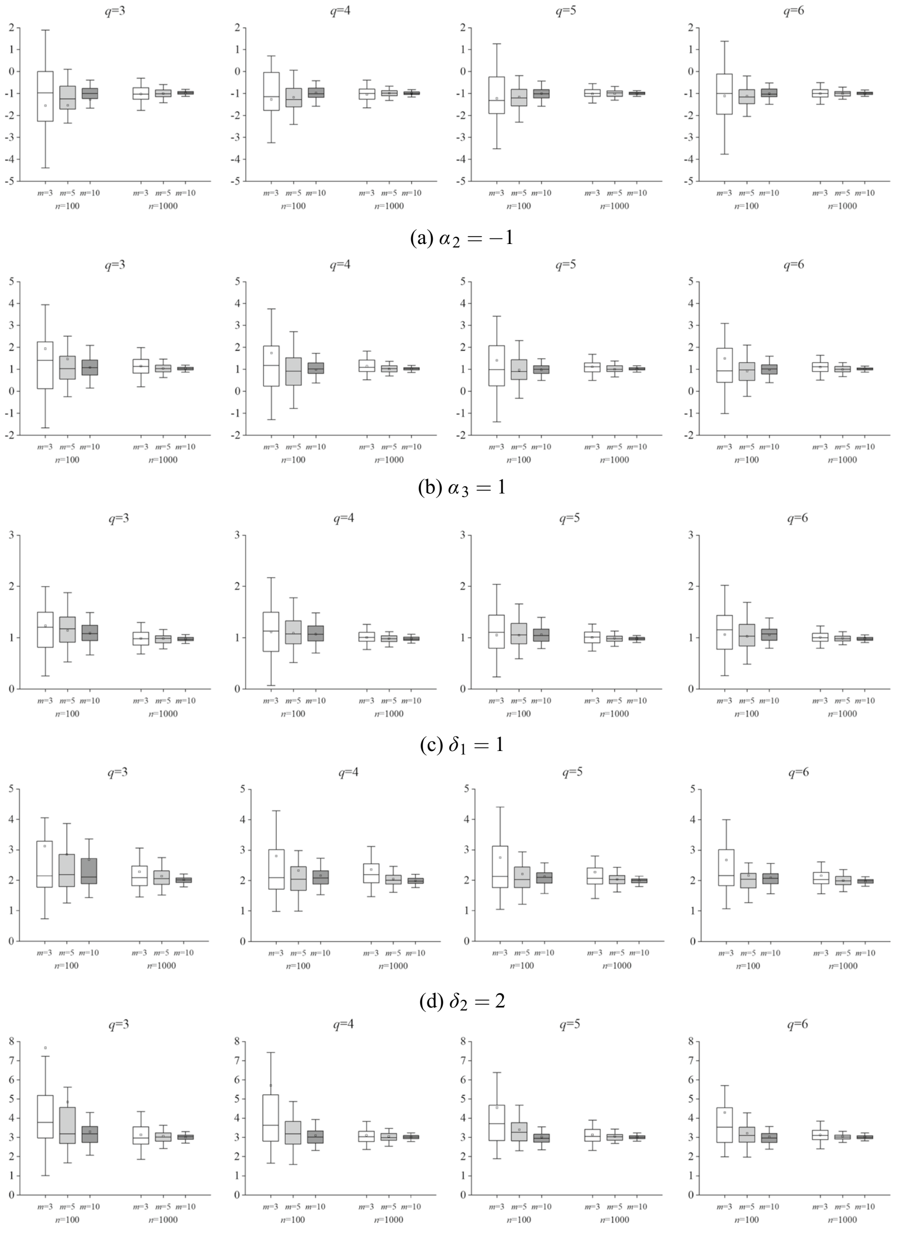
Figures 1 and 2 provide a summary of the results of Scenario 2 for $R = 3$ for the models (4) and (3). We observe that the performance is remarkably similar to that of Scenario 1. In other words, the estimates of parameters in Model (4) are closer to their true values than those in Model (3), and the performance improves as the number of response categories, q increases. The results of our simulation study

Table 1 Scenario 1: The average absolute bias and standard error obtained for each parameter over 2000 simulations for models formulated in Eqs. (3) and (4)

	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
Model 3									
	$\alpha_2 = -1$	0.52	0.05	0.54	0.04	0.55	0.03	0.58	0.03
$n = 100$	$\alpha_3 = -2$	2.28	0.16	2.46	0.18	2.25	0.14	2.08	0.07
$R = 5, m = 10$	$\alpha_4 = 1$	0.65	0.05	0.64	0.03	0.69	0.03	0.72	0.02
	$\alpha_5 = 2$	3.47	0.19	3.16	0.14	2.98	0.12	2.90	0.11
	$\alpha_2 = -1$	0.46	0.01	0.52	0.01	0.57	0.01	0.60	0.01
$n = 1000$	$\alpha_3 = -2$	2.21	0.02	2.19	0.01	2.23	0.01	2.26	0.01
$R = 5, m = 10$	$\alpha_4 = 1$	0.49	0.01	0.58	0.01	0.65	0.01	0.68	0.01
	$\alpha_5 = 2$	2.63	0.02	2.70	0.01	2.79	0.01	2.83	0.01
Model 4									
	$\alpha_2 = -1$	0.26	0.03	0.26	0.03	0.27	0.03	0.25	0.03
$n = 100$	$\alpha_3 = -2$	0.59	0.07	0.46	0.05	0.39	0.04	0.38	0.04
$R = 5, m = 10$	$\alpha_4 = 1$	0.37	0.04	0.30	0.03	0.30	0.03	0.29	0.03
	$\alpha_5 = 2$	0.51	0.07	0.34	0.04	0.33	0.04	0.29	0.03
	$\delta_1 = -1$	0.15	0.02	0.15	0.01	0.15	0.02	0.16	0.01
	$\delta_2 = 2$	0.21	0.02	0.18	0.02	0.17	0.02	0.14	0.01
	$\delta_3 = -3$	0.69	0.08	0.58	0.07	0.47	0.06	0.38	0.05
	$\delta_4 = 4$	0.70	0.08	0.49	0.05	0.41	0.05	0.38	0.04
	$\delta_5 = -5$	1.55	0.20	1.06	0.13	0.95	0.13	0.81	0.11
	$\alpha_2 = -1$	0.09	0.01	0.08	0.01	0.08	0.01	0.08	0.01
$n = 1000$	$\alpha_3 = -2$	0.09	0.01	0.09	0.01	0.08	0.01	0.08	0.01
$R = 5, m = 10$	$\alpha_4 = 1$	0.11	0.01	0.10	0.01	0.10	0.01	0.09	0.01
	$\alpha_5 = 2$	0.13	0.01	0.10	0.01	0.09	0.01	0.09	0.01
	$\delta_1 = -1$	0.04	0.01	0.03	0.01	0.03	0.01	0.03	0.01
	$\delta_2 = 2$	0.09	0.01	0.07	0.01	0.06	0.01	0.06	0.01
	$\delta_3 = -3$	0.12	0.01	0.10	0.01	0.09	0.01	0.08	0.01
	$\delta_4 = 4$	0.19	0.01	0.16	0.01	0.16	0.01	0.13	0.01
	$\delta_5 = -5$	0.25	0.01	0.19	0.01	0.18	0.01	0.16	0.01

indicate that the clustering procedure described in this article is able to recover the true parameter values in all tested instances. In addition, the other results for this scenario are shown in Tables 8, 9, 10, and 11 and Figs. 8, and 9 in Appendix A.

Finally, Fig. 3 shows the average ARI, 1-NVI, and 1-NID over all the replicates and different numbers of rows, by Scenario, R, and q. These measures compare the similarity of resulting row clustering structures in the true models with the fitted models. It can be seen that, for Scenario 1, the mean of similarity measures based on the ARI is 0.65 between the true clustering memberships and the predicted memberships when data were fitted by Model (4) with $R = 3$, $q = 3$. Similarly, the mean of similarity measures based on the ARI is 0.69 when $R = 3$, $q = 5$. Thus, the measure



(a) $\alpha_2 = -1$

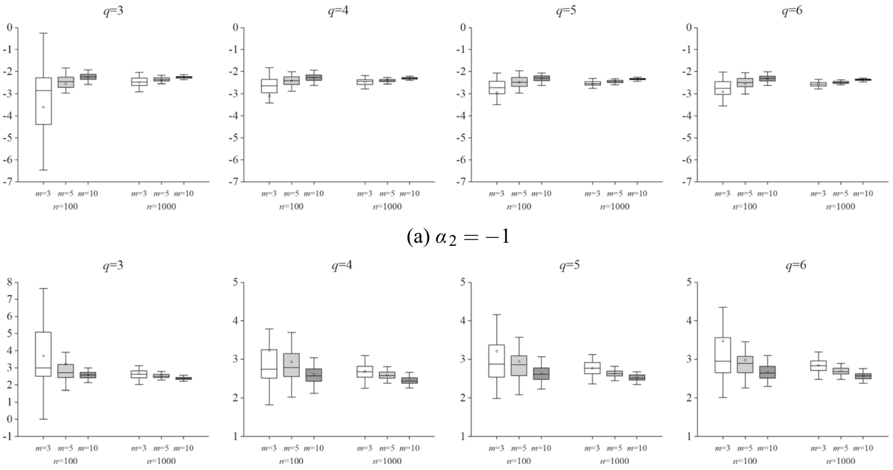
(b) $\alpha_3 = 1$

(c) $\delta_1 = 1$

(d) $\delta_2 = 2$

(e) $\delta_3 = 3$

Fig. 1 Boxplots for Scenario 2, representing the estimated distribution of each parameter when the fitted model is Model 4 and R = 3



(a) $\alpha_2 = -1$

(b) $\alpha_3 = 1$

Fig. 2 Boxplots for Scenario 2, representing the estimated distribution of each parameter when the fitted model is Model 3 and R = 3

increases with increasing q, most likely because the data with $q = 5$ contain more information about the response. We observed equivalent results for the other two measures: 1-NVI and 1-NID. For Scenario 2, all three measures (ARI, 1-NVI and 1-NID) show equivalent results to Scenario 1 but the indices are smaller. For example, the ARI is 0.46 when $R = 3$, $q = 3$ and the ARI is 0.49 when $R = 3$, $q = 5$. All three clustering measures have smaller values than the ones in Scenario 1. We conclude from this that if some covariate effects are positive and others are negative (Scenario 1), then it is easier to detect the correct clustering structure than if the covariate effects are all in the same direction (Scenario 2).

5 Application

We applied the models proposed in this article to the *arthritis clinical trial* data set (Lipsitz et al. 1996), which compares the drug auranofin and placebo therapy for the treatment of rheumatoid arthritis. The data set is obtained from the **R** package *multgee* (Touloumis 2015). The response variable is the patient's self-assessment of arthritis, which is measured on a five-level ordinal response scale, from very poor (1) to very good (5). A total of 302 eligible patients were in the original data set but only 289 patients completed a rheumatoid self-assessment questionnaire at all three follow-up times (first, third and fifth month of treatment). We used those 289 patients with completed questionnaires to analyze in our example. The data can be represented by a 289 × 3 matrix **Y**. The covariates we include in our model are gender (1=female and 0=male), age (in years), and treatment (1=placebo and 0=drug). In this application, the covariate-dependent clustering could help to identify subsets of patients with similar covariate information patterns. This insight would be

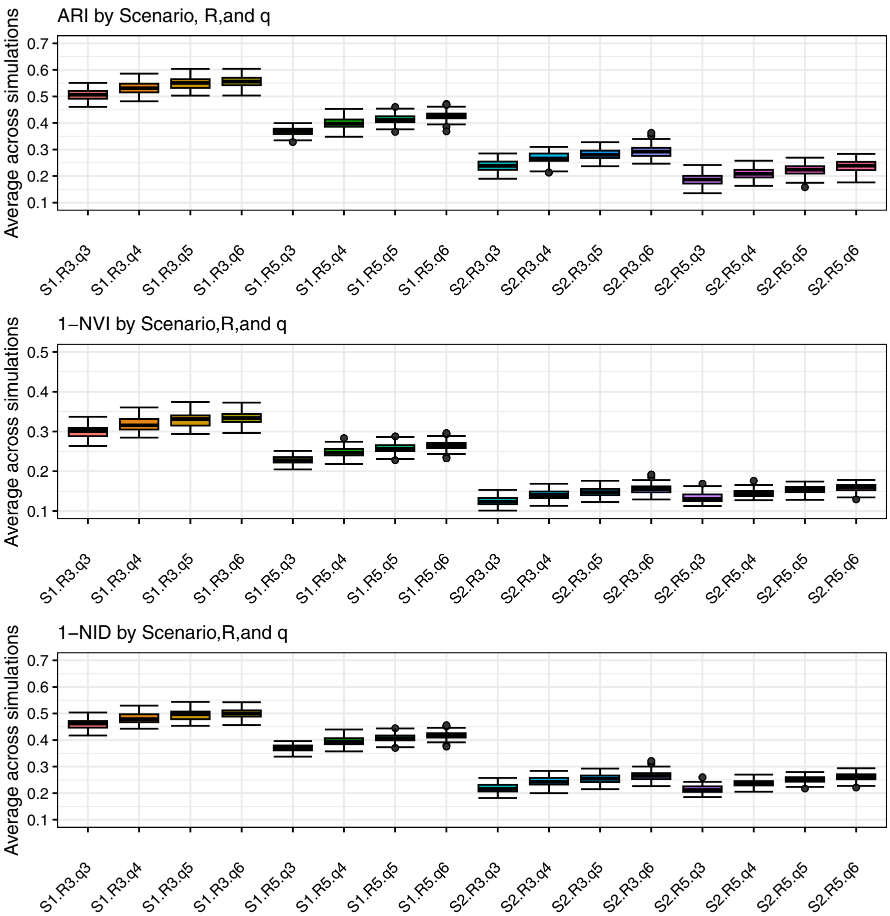


Fig. 3 Boxplot: Clustering structure comparison for simulations between Scenario 1 and Scenario 2 using ARI, 1-NVI and 1-NID. Values in the vertical axis indicate averages across 2000 simulations. The labels in y-axis are in the format "Sx.Ry.qz" being x the number of scenario (1, 2), y the number of clusters ($R = 3.5$), and z the number of ordinal categories ($q = 3, 4, 5, 6$)

important because it would provide a flexible approach for identifying potential heterogeneous gender, age, and auranofin treatment effects on the arthritis scores. For instance, if the elderly experience more symptoms and, consequently, tend to be more pessimistic about their arthritis status, our proposed model would allow us to distinguish subsets of older people that tend to report higher/lower arthritis scores. However, we note that this is only an example and we do not advocate the clinical relevance of the covariate-dependent clustering model. In real settings, clinicians and the statisticians together should decide which model, i.e. no clustering, clustering with covariates, or clustering without covariates, is more relevant to answer their research questions. After fitting the models without covariates (3) and with covariates (4), with different number of row clusters, we compared them using

Table 2 Results of row clustering models fitted to the arthritis data set

Model	R	Number of parameter	Log-like	AIC	BIC
$\mu_k - \alpha_r$	2	6	−1096.99	2205.99	2234.58
	3	8	−1077.73	2171.46	2209.59
	4	10	−1067.20	**2154.40**	**2202.05**
	5	12	−1067.20	2158.40	2215.58
$\mu_k - (\alpha_r + x_i\delta_r)$					
$x =$ age	2	8	−1138.18	2292.37	2330.49
	3	11	−1071.88	2165.75	**2218.17**
	4	14	−1065.18	2158.37	2225.08
	5	17	−1060.84	**2155.68**	2236.68
$x =$ treatment	2	8	−1082.28	2180.57	2218.69
	3	11	−1067.93	2157.87	2210.28
	4	14	−1057.70	**2143.40**	**2210.11**
	5	17	−1056.23	2146.46	2227.46
$x =$ gender	2	8	−1096.89	2209.77	2247.89
	3	11	−1079.51	2181.02	2233.44
	4	14	−1066.92	**2161.84**	**2228.55**
	5	17	−1066.37	2166.74	2247.74
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$					
$x_1 =$ age,	2	10	−1072.54	2165.07	**2212.72**
$x_2 =$ treatment	3	14	−1059.23	2146.46	2213.17
	4	18	−1050.39	**2136.78**	2222.55
	5	22	−1048.53	2141.05	2245.88
$x_1 =$ age,	2	10	−1085.83	2191.67	2239.32
$x_2 =$ gender	3	14	−1068.97	2165.95	**2232.66**
	4	18	−1061.29	**2158.58**	2244.35

Table 2 (continued)

Model	R	Number of parameter	Log-like	AIC	BIC
	5	22	−1059.26	2162.52	2267.35
$x_1 = \text{treatment},$	2	10	−1081.82	2183.64	2231.29
$x_2 = \text{gender}$	3	14	−1065.99	2159.99	**2226.71**
	4	18	−1056.73	**2149.45**	2235.22
	5	22	−1055.06	2154.13	2258.96
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r} + x_{i3}\delta_{3r})$					
$x_1 = \text{age},$	2	12	−1071.60	2167.21	**2224.39**
$x_2 = \text{treatment},$	3	17	−1060.50	2155.00	2236.01
$x_3 = \text{gender}$	4	22	−1050.35	**2144.71**	2249.54
	5	27	−1052.14	2158.35	2287.00

The best model in each group of models (no covariates, one, two, or three covariates), based on AIC and BIC, is shown in bold

Table 3 Estimated parameters of two models, the first with no covariates and the second with the covariates age and treatment

	Model without covariates (3)	Model with covariates (4)		
R	α_r	α_r	δ_{1r} (age)	δ_{2r} (treatment)
1	4.20 (0.17)	3.64 (0.12)	0.56 (0.15)	0.23 (0.14)
2	1.26 (0.25)	1.20 (0.15)	0.02 (0.25)	−0.84 (0.22)
3	−1.41(0.14)	−1.50 (0.24)	−0.22 (0.12)	−0.82 (0.25)
4	−4.04 (0.13)	−3.34 (0.18)	0.58 (0.13)	−2.00 (0.17)

the information criteria AIC (Akaike 1973) and BIC (Schwarz 1978) (see results in Table 2). AIC indicates that the best model is the version of the row clustering model including age and treatment covariates ($\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$) with $R = 4$ row clusters (AIC = 2136.78), which is better than its counterpart in the model without covariates (AIC=2154.40). However, BIC shows that the model without covariates ($\mu_k - \alpha_r$) and $R = 4$ is the best model (BIC=2202.05). A possible reason is that BIC penalizes higher numbers of parameters more strongly than AIC does, leading to a preference for more parsimonious models. On the other hand, based on our experience working with practitioners and researchers from other areas, we have chosen to use AIC as a standard measure for model selection. Nevertheless, we also acknowledge the importance of BIC in providing more parsimonious models, and we have included the results from BIC in our analyses to ensure a comprehensive evaluation of model performance.

Table 3 shows the estimated parameters of the two models. The row clustering model without covariates (3) separates the patients into four clusters (sorted by the best to the worst self-assessment scores). The first cluster has the strongest patient feelings effect, ($\alpha_1 = 4.20$), followed by cluster 2 ($\alpha_2 = 1.26$), cluster 3 ($\alpha_3 = -1.41$), and cluster 4 ($\alpha_4 = -4.04$), suggesting that patients in cluster 1 have the best feeling about their current state of arthritis among all the clusters. When we add the age and treatment covariates into the clustering model, the parameters $\{\delta_{1r}\}$ and $\{\delta_{2r}\}$ indicate the age and treatment effects within the clusters. For instance, the auranofin treatment did not show improvement for patients in cluster 1 ($\delta_{21} = 0.23$), but the treatment did show improvement, to differing degrees, for patients in clusters 2, 3, and 4 ($\delta_{22} = -0.84$, $\delta_{23} = -0.82$ and $\delta_{24} = -2.00$). Moreover, the older patients in cluster 3 ($\delta_{13} = -0.22$) were likely to have a worse feeling about their current arthritis status than older patients in other clusters. Therefore, the clustering model without covariates (3) allows us to describe the overall patterns of patient feelings and once we add the covariates (4), we could also identify the subgroups of patients with similar covariate patterns.

Additionally, Table 12 (see results in Appendix A) shows the results of the comparison of clustering structure agreement of the selected models with and without covariates by using the information theoretic criteria ARI, 1-NVI and 1-NID. The results assume each patient has been allocated to the cluster for which they have the highest posterior probability of membership.

The comparison of clustering structure agreement (measured by ARI, 1-NVI, and 1-NDI) between the best model (Model (4) with $R = 4$, including age and treatment covariates) and its counterpart without covariates (Model (3)) revealed distinct differences. The values of the three measures were 0.66 (ARI), 0.47 (1-NVI), and 0.64 (1-NDI), indicating that Model (3) and Model (4) resulted in different clustering structures. This was further confirmed by examining the detailed estimated memberships for individuals in Table 13 (see results in Appendix A). For example, when data were fitted by Model (4), nine patients (1, 62, 119, 124, 131, 223, 239, 243, and 266) originally assigned to cluster 1 were re-allocated to cluster 2. Similarly, two patients (79 and 238) from cluster 2 were re-allocated to cluster 3, and ten patients (63, 125, 153, 192, 215, 217, 219, 245, 267, and 285) from cluster 3 were re-allocated to cluster 4. These findings highlight the substantial impact of including covariates in Model (4) on the clustering structures, underscoring the importance of considering covariate effects in the analysis.

Finally, Table 14 in Appendix C presents a comparison of the average age for different combinations of clusters and treatment (placebo or drug) using two models: one without covariates ($\mu_k - \alpha_r$) with $R = 4$ clusters, and the other incorporating the covariates age and treatment ($\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r}$), where x_1 represents age and x_2 represents treatment) also with $R = 4$ clusters. Notable differences in mean age were observed within specific groups. For instance, individuals in group 1 receiving the drug treatment exhibited an increase in mean age from 35 to 41 when covariates were incorporated. Conversely, a similar discrepancy in mean age, but in the opposite direction, was observed for individuals in group 2 receiving the placebo (54.5 vs 50.8). This comparison highlights the added insights provided by including covariates, shedding light on the relationship between variables and their impact on mean age within specific groups and treatment categories. These findings underscore the importance of considering covariates in understanding population characteristics and their potential influence on outcomes.

6 Discussion

This paper uses finite mixture models to model the case of ordinal data using the proportional odds model, including covariates in the linear predictor of the model. We used the proportional odds model to capture the inherent natural order of the responses.

We set up a simulation study to explore the reliability of the models with covariates across a range of cases. We considered two scenarios by varying the covariate effects from mixed directions to all in the same direction. For all cases, the estimates of the parameters are close to their true values and we observed that the value of the standard errors decreases as the number of ordinal categories increases. The standard errors also decrease with increasing sample size n. Moreover, we compared the similarity of the true model and the fitted model for both scenarios based on ARI, 1-NVI, and 1-NID indices. The row clustering structure with mixed direction covariate effects showed better performance than the one with all positive covariate effects.

We also illustrated our approach with the well-known "*arthritis clinical trial*" data set. The results of this application indicated that the best model according to AIC was the row clustering model with $R = 4$ including age and treatment covariates. However, we remark that AIC is a standard procedure and we consider that subject-matter experts in the matter and statisticians together should decide which model (adding covariates or not) is more relevant to answer the research questions. The patients were clustered according to their similar pattern of responses and the effect of the covariates. That is, all four clusters have different age and treatment effects, which changes the interpretation of the clustering structure when the covariates were not taken into account. In this case, we could identify individuals in each of the four groups based on their self-assessment scales and how the age and treatment are associated with these groups.

It is important to note that our proposed model is based on the cumulative version of the proportional odds model, which applies the proportional odds assumption. This assumption must be assessed in different ways: (1) examining graphical diagnostics, e.g. plotting the cumulative logit probabilities against the covariates can reveal any systematic departures from parallelism, which is an indication of violations of the proportional odds assumption; (2) performing formal statistical tests, e.g. the Brant Wald test (1990); and (3) performing model diagnostics, e.g. examining the residuals for patterns, such as non-linearity or heteroscedasticity, which can provide evidence of violations of the assumptions. Additionally, we acknowledge that this represents a simplified approach that assumes a uniform covariate effect, which may not always be valid in all cases and may not capture the true complexity of the relationship between covariates and response variables in all cases. Further research is needed to explore more flexible models that can account for varying covariate effects on different response variables in different situations.

Deciding whether a variable should be considered a response or a covariate is a crucial step in statistical analyses, including mixture-based clustering. By necessity, statistical analyses distinguish between response (dependent) variables and explanatory (independent) variables (Agresti 2014). In making this decision, it is essential to consider the research questions, theoretical considerations, and subject-matter knowledge about the variables under study. Researchers need to carefully assess whether a variable is of primary interest in the study to answer the research questions, or if it plays a supporting role because of its association with the variable of interest. Additionally, exploratory data analysis techniques, such as visualizations and correlation analyses, can provide valuable insights into the relationships between variables. By considering these factors and utilizing appropriate statistical techniques, researchers can make informed decisions regarding the allocation of variables as responses or covariates. In this study, we follow the general convention of treating variables as responses or covariates based on research interest. For instance, in our application, the main interest is the effect of a drug on the arthritis status. However, if a researcher is investigating the potential average age of the patient according to their arthritis status, the age variable would likely be considered a response variable. We acknowledge that this allocation of variables may vary depending on the specific research context, and researchers should adapt their approach accordingly.

Our approach assumes that all available variables are used in the modeling procedure. However, in many situations, considering all the variables unnecessarily increases the model complexity. Moreover, some variables may not possess any clustering information and are of no use in the detection of the group structure. Rather, they could be detrimental to the clustering. Likewise, the case where all the variables contain clustering information can also be problematic. Along with the increasing number of dimensions comes the curse of dimensionality, and including superfluous variables in the model leads to identifiability problems and over-parameterization (Bouveyron and Brunet-Saumard 2014). Therefore, resorting to variable selection techniques can facilitate model fitting, ease the interpretation of the results and lead to data classification of better quality. Even in situations of moderate or low dimensionality, reducing the set of variables employed in the clustering process can be beneficial (Fowlkes et al. 1988; Raftery and Dean 2006; Andrews and McNicholas 2014). How the variable selection algorithm interacts with the model fitting process defines the overall approach to the problem. For a general learning task, the principal distinction is in whether the selection is carried out separately or jointly to the learning procedure (John et al. 1994; Dash and Liu 1997; Dy and Brodley 2004). Thus, the application of information criteria such as AIC or BIC would be a direct way to perform variable selection in our approach. There would be other alternatives as model-based selection methods, such as stepwise selection and LASSO. Additionally, domain knowledge or subject-matter expertise can also guide the variable selection process by considering the relevance of covariates based on their theoretical importance or prior knowledge.

We performed a robustness analysis to assess the impact of outliers by introducing 3% outliers in the numerical variable age and re-fitting the models incorporating this covariate. We have included these results in Table 5 of Appendix E. Interestingly, our analysis consistently revealed the emergence of an additional group, according to AIC, resulting in a total of five groups, in contrast to the framework without outliers. Importantly, this newly identified group aligned with the rows containing the artificially introduced outliers. These findings demonstrate the sensitivity of our proposed method to outliers and its ability to capture their influence on the clustering structure. However, it is essential to note that this analysis serves as an illustrative example, and a more comprehensive robustness analysis is warranted in future research. Therefore, we view this as a potential avenue for future investigations.

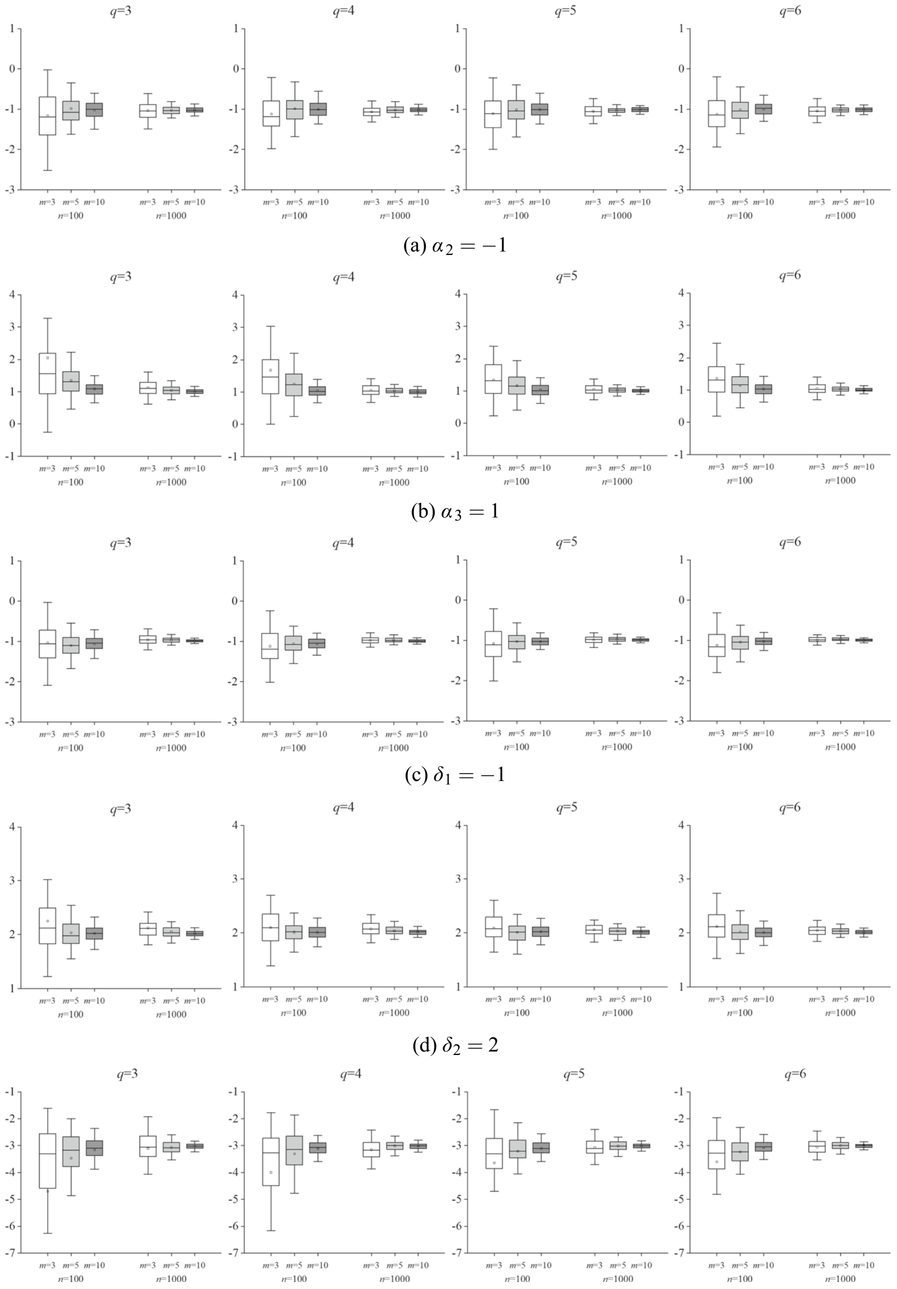
We compared the best model according to AIC with the Partitioning Around Medoids (PAM) method using the Gower dissimilarity measure to assess the equivalence in terms of the number of clusters and the cluster structure of covariate values. The results of this comparison can be found in Table 15 of Appendix D. Interestingly, this comparison revealed consistent results in the number of groups ($R = 4$), while exhibiting slight differences in both the cluster structures and covariate values. Notably, the incorporation of covariates age and treatment in the model $(\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r}))$ resulted in a lower mean age for individuals in group 1 receiving drug treatment, compared to the results obtained with the PAM method. A more comprehensive comparison of clustering methods would be an intriguing avenue for future research.

This article demonstrated that including available covariates in the fitting process of the mixture-based approaches for ordinal responses improves insights into the main characteristics of the clusters. The same idea could be implemented for different types of models focused on ordinal responses, such as the ordered stereotype and the adjacent-categories logit models. For future research, we will plan to extend the model shown in this article in order to cluster rows and columns simultaneously (a.k.a. co-clustering or biclustering), which is a natural extension and can give more insights of the clustering structure of the data sets. Additionally, another natural and challenging extension to consider would be to incorporate rows and columns covariates to our current approach, capturing their potential interactions. One potential approach might be using a multi-level modeling framework or fitting two separate mixture models. Another interesting avenue to explore would be the potential application of our proposed procedure for data imputation in case of missing data. It could be extended to impute both ordinal responses and covariate values simultaneously, leveraging the estimated mixture models and capturing non-linear relationships and interactions between variables. The uncertainty-aware imputation approach using the EM algorithm could provide more realistic and robust imputed values. However, further research and validation would be needed to evaluate the performance of our proposed procedure as a data imputation method, in comparison to existing techniques, in various settings and data scenarios. Additionally, as a future work, we plan to conduct additional comparisons with other existing methods to further evaluate the performance of our proposed method and provide a more comprehensive analysis. Finally, this research has considered the case where responses in each column have the same number of ordinal response levels. This could be varied but may require a separate set of parameters $\{\mu_{jk}\}$ and $\{\phi_{jk}\}$. The simulation and model fitting code in **R** is available on Github at https://github.com/vuw-clustering/clustering-covariates.

Appendices

A Simulation results

See Figs. 4, 5, 6, 7, 8, and 9 and Tables 4, 5, 6, 7, 8, 9, 10, and 11.



(a) $\alpha_2 = -1$

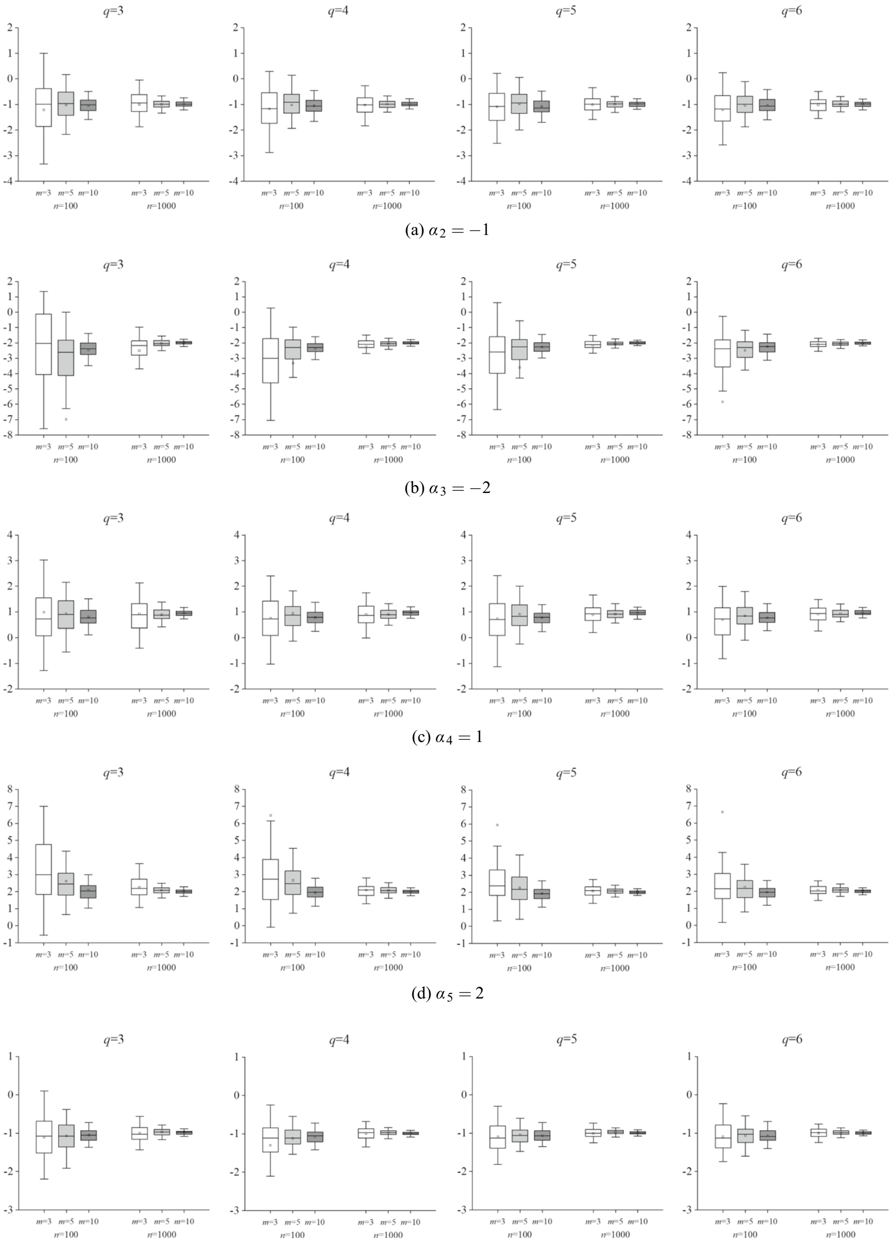
(b) $\alpha_3 = 1$

(c) $\delta_1 = -1$

(d) $\delta_2 = 2$

(e) $\delta_3 = -3$

Fig. 4 Boxplots for Scenario 1 representing the estimated distribution of each parameter when the fitted model is Model 4 and $R = 3$



(a) $\alpha_2 = -1$

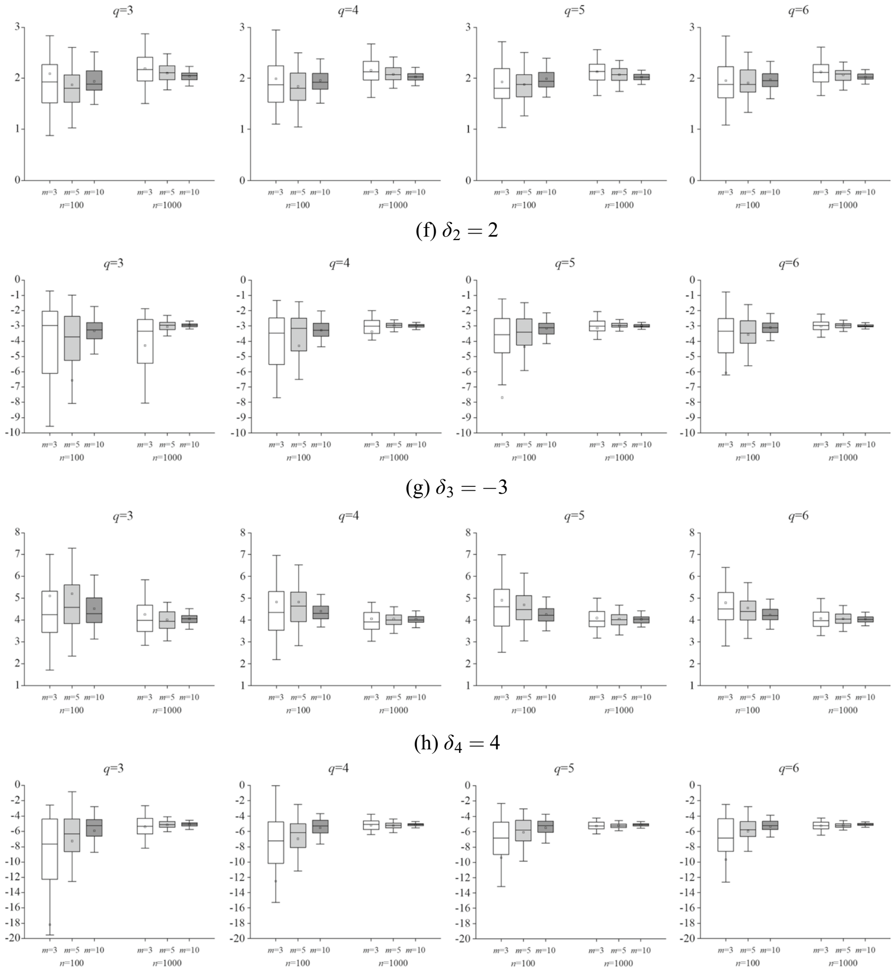
(b) $\alpha_3 = -2$

(c) $\alpha_4 = 1$

(d) $\alpha_5 = 2$

(e) $\delta_1 = -1$

Fig. 5 Boxplots for Scenario 1, representing the estimated distribution of each parameter when the fitted model is Model 4 and $R = 5$



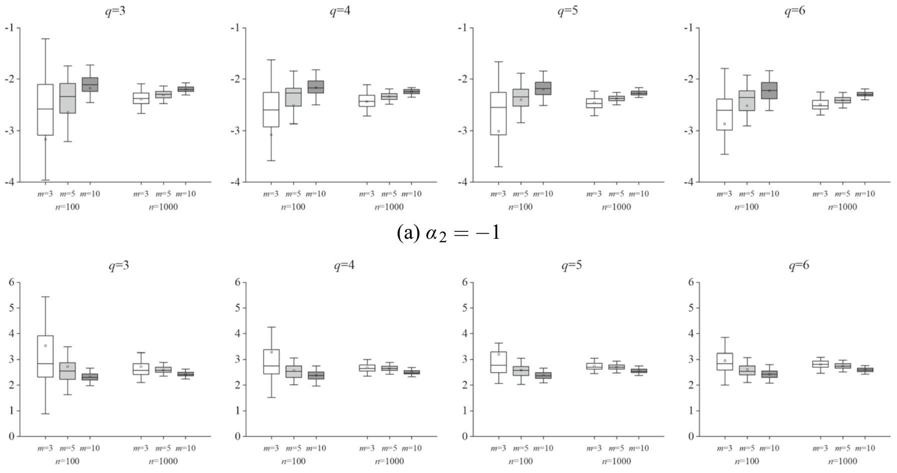
(f) $\delta_2 = 2$

(g) $\delta_3 = -3$

(h) $\delta_4 = 4$

(i) $\delta_5 = -5$

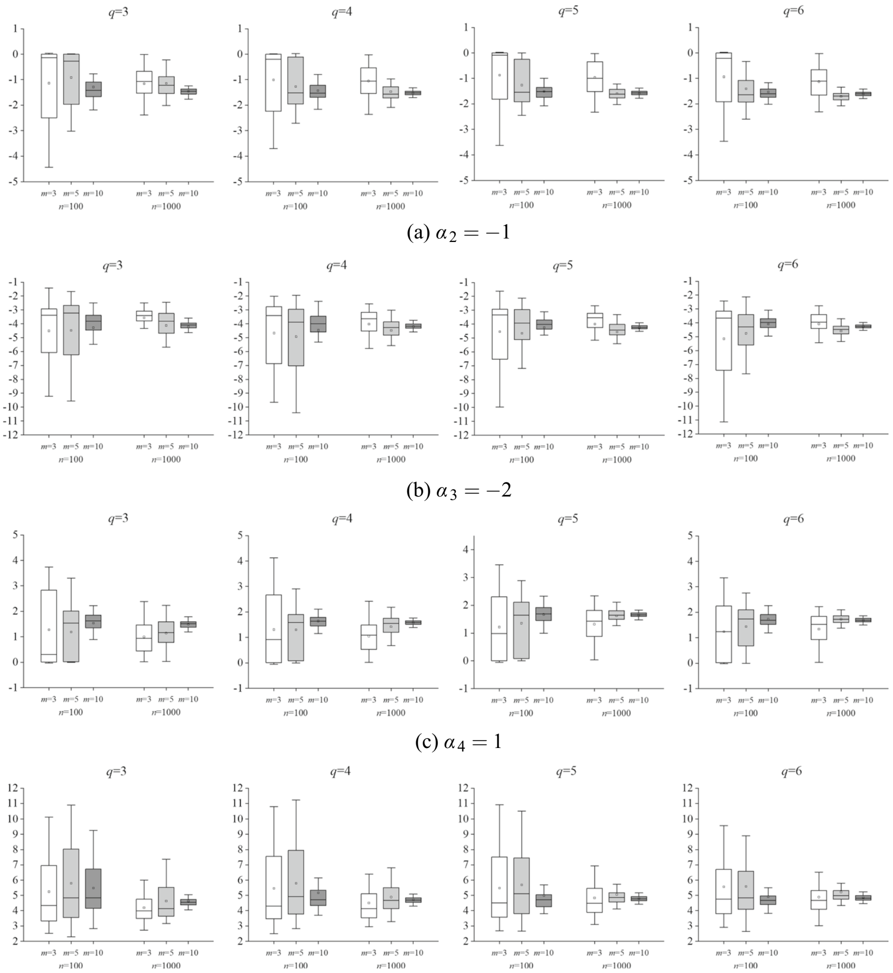
Fig. 5 (continued)



(a) $\alpha_2 = -1$

(b) $\alpha_3 = 1$

Fig. 6 Boxplots for Scenario 1, representing the estimated distribution of each parameter when the fitted model is Model 4 and $R = 5$



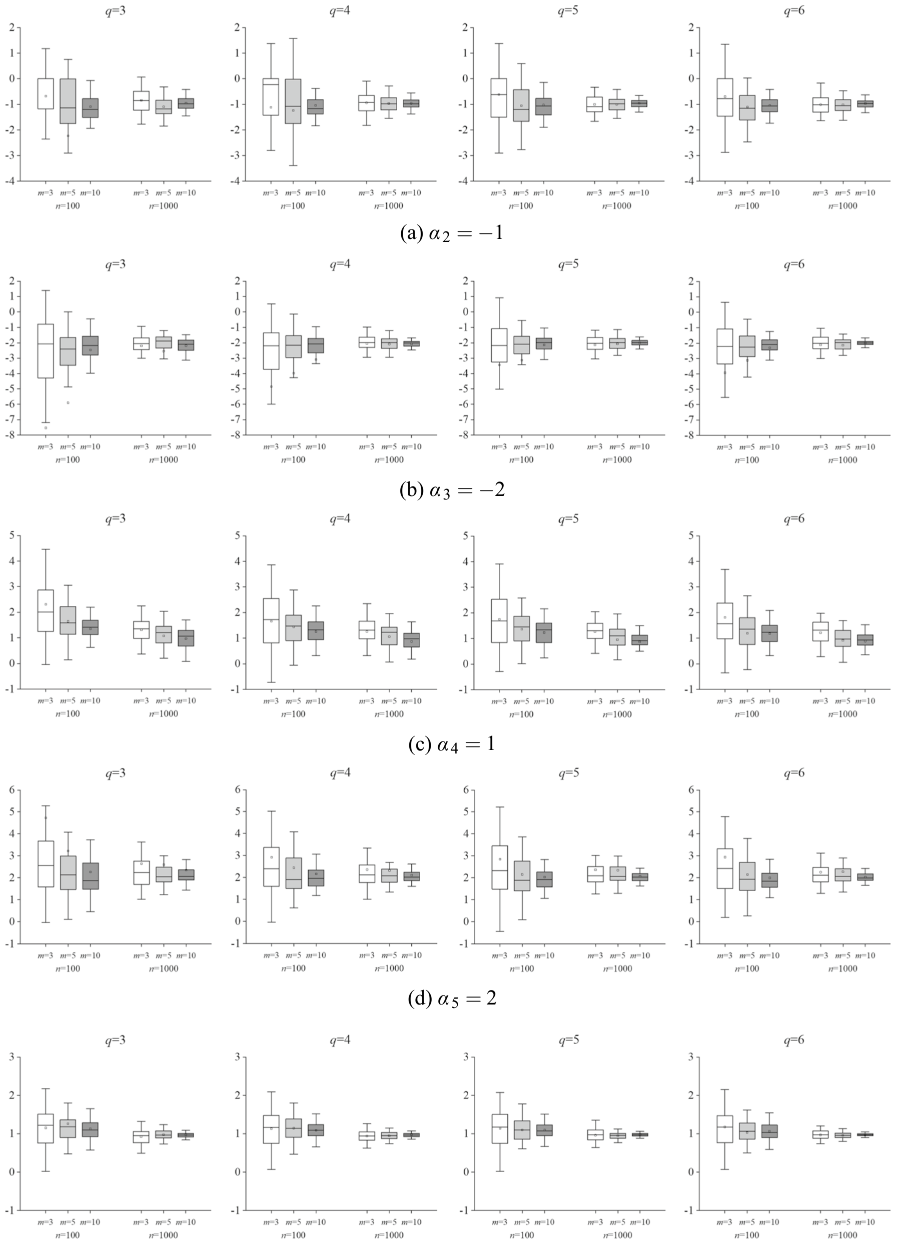
(a) $\alpha_2 = -1$

(b) $\alpha_3 = -2$

(c) $\alpha_4 = 1$

(d) $\alpha_5 = 2$

Fig. 7 Boxplots for Scenario 1, representing the estimated distribution of each parameter when the fitted model is Model 3 and $R = 5$



(a) $\alpha_2 = -1$

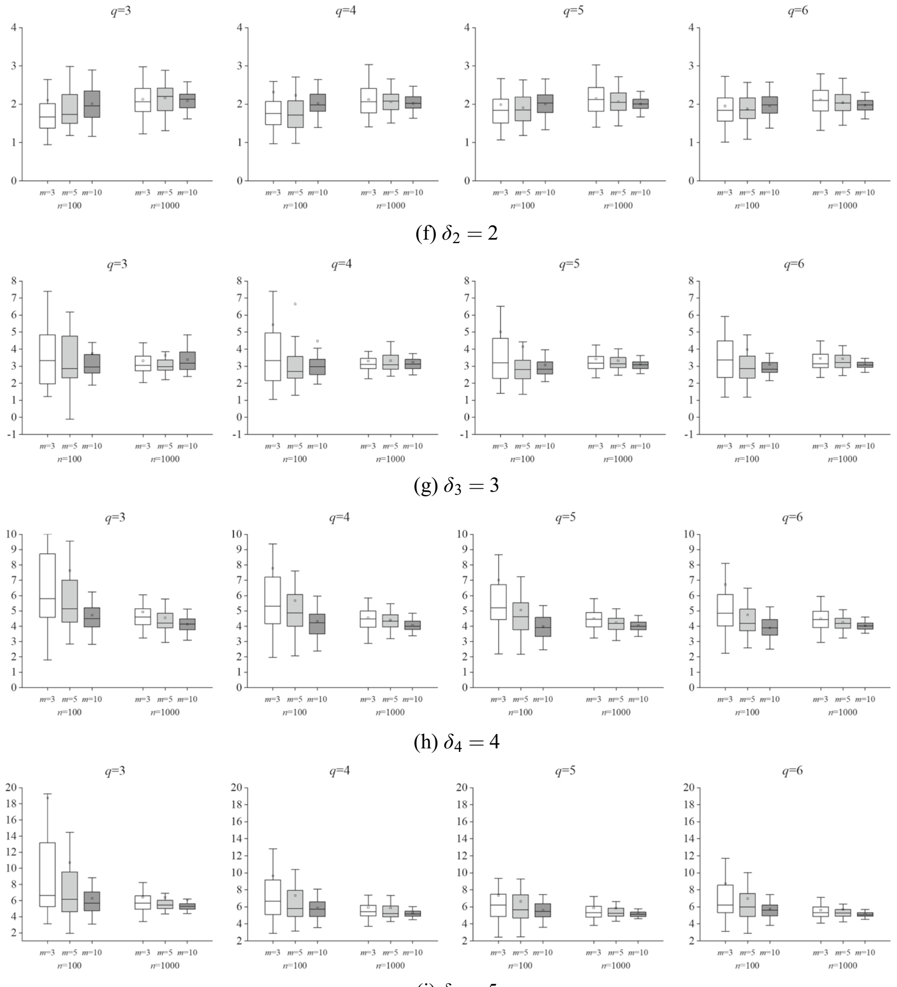
(b) $\alpha_3 = -2$

(c) $\alpha_4 = 1$

(d) $\alpha_5 = 2$

(e) $\delta_1 = 1$

Fig. 8 Boxplots for Scenario 2, representing the estimated distribution of each parameter when the fitted model is Model 4 and $R = 5$



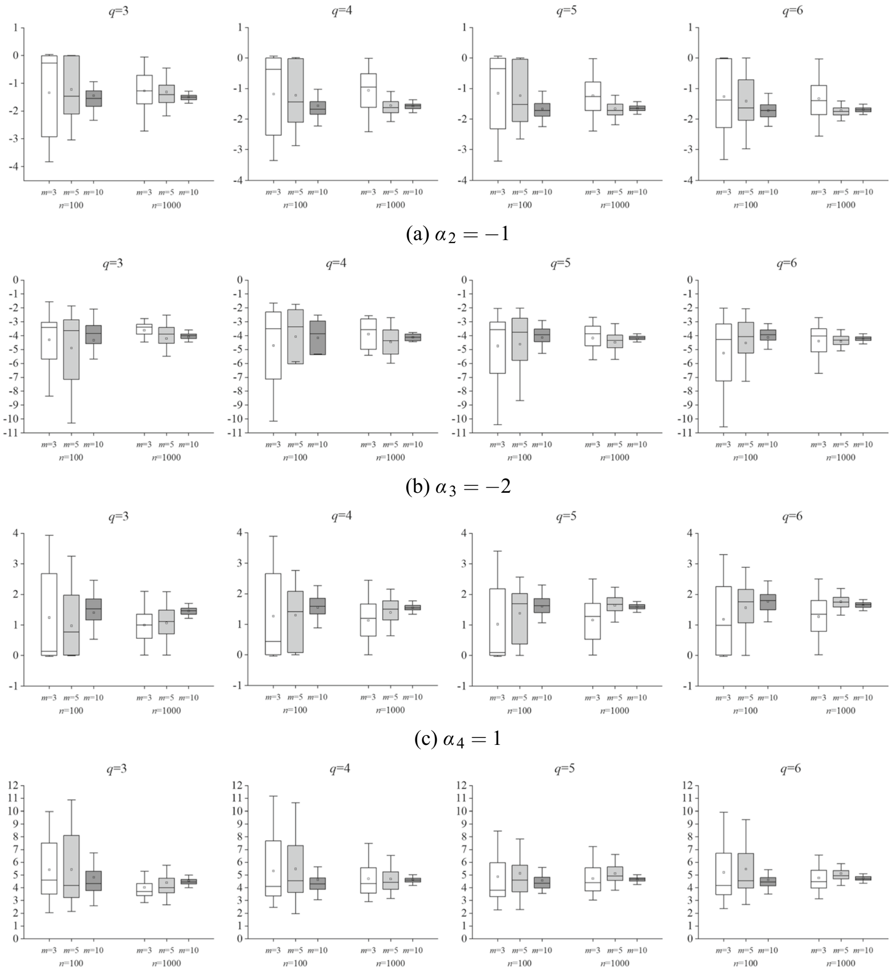
(f) $\delta_2 = 2$

(g) $\delta_3 = 3$

(h) $\delta_4 = 4$

(i) $\delta_5 = 5$

Fig. 8 (continued)



(a) $\alpha_2 = -1$

(b) $\alpha_3 = -2$

(c) $\alpha_4 = 1$

(d) $\alpha_5 = 2$

Fig. 9 Boxplots for Scenario 2, representing the estimated distribution of each parameter when the fitted model is Model 3 and $R = 5$

Table 4 Scenario 1: The average absolute bias and standard error obtained for each parameter over 2000 simulations when the fitted model is Model 4 (n = 100)

n = 100									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
$m = 3$									
3	$\alpha_2 = -1$	0.63	0.09	0.48	0.07	0.44	0.06	0.40	0.05
	$\alpha_3 = 1$	1.40	0.23	1.02	0.18	0.67	0.08	0.71	0.10
	$\delta_1 = -1$	0.46	0.06	0.39	0.04	0.36	0.04	0.36	0.04
	$\delta_2 = 2$	0.48	0.07	0.29	0.03	0.25	0.03	0.26	0.03
	$\delta_3 = -3$	2.11	0.62	1.33	0.22	0.98	0.17	0.88	0.14
5	$\alpha_2 = -1$	0.96	0.13	0.86	0.12	0.65	0.08	0.63	0.08
	$\alpha_3 = -2$	11.68	8.49	11.80	4.61	7.17	3.42	4.43	2.75
	$\alpha_4 = 1$	1.12	0.18	0.96	0.13	0.82	0.10	0.74	0.09
	$\alpha_5 = 2$	10.22	4.20	5.09	2.68	4.45	2.98	5.31	3.99
	$\delta_1 = -1$	0.54	0.07	0.62	0.15	0.38	0.04	0.40	0.05
	$\delta_2 = 2$	0.59	0.09	0.47	0.06	0.46	0.06	0.42	0.05
	$\delta_3 = -3$	15.22	5.95	8.31	3.24	5.19	2.35	3.66	1.92
	$\delta_4 = 4$	1.80	0.40	1.30	0.19	1.24	0.16	1.06	0.13
	$\delta_5 = -5$	13.87	4.41	8.27	3.00	5.01	1.63	5.49	2.35
$m = 5$									
3	$\alpha_2 = -1$	0.34	0.04	0.31	0.04	0.26	0.03	0.25	0.03
	$\alpha_3 = 1$	0.50	0.05	0.50	0.06	0.36	0.04	0.33	0.04
	$\delta_1 = -1$	0.25	0.03	0.22	0.02	0.21	0.03	0.20	0.03
	$\delta_2 = 2$	0.24	0.03	0.18	0.02	0.16	0.02	0.16	0.02
	$\delta_3 = -3$	0.84	0.12	0.69	0.10	0.49	0.07	0.49	0.06
5	$\alpha_2 = -1$	0.55	0.07	0.50	0.07	0.50	0.07	0.42	0.05
	$\alpha_3 = -2$	5.42	1.73	1.65	0.69	2.00	1.14	0.73	0.10
	$\alpha_4 = 1$	0.69	0.09	0.60	0.09	0.55	0.07	0.46	0.06
	$\alpha_5 = 2$	0.96	0.12	0.97	0.13	0.72	0.08	0.63	0.08
	$\delta_1 = -1$	0.35	0.04	0.29	0.04	0.25	0.04	0.23	0.03
	$\delta_2 = 2$	0.40	0.05	0.34	0.04	0.27	0.03	0.26	0.03
	$\delta_3 = -3$	4.16	1.19	1.82	0.50	1.86	0.83	1.01	0.12
	$\delta_4 = 4$	1.56	0.23	1.07	0.13	0.90	0.12	0.73	0.10
	$\delta_5 = -5$	2.99	0.40	2.40	0.29	1.68	0.20	1.49	0.18
$m = 10$									
3	$\alpha_2 = -1$	0.18	0.02	0.16	0.02	0.14	0.01	0.14	0.02
	$\alpha_3 = 1$	0.21	0.02	0.16	0.02	0.17	0.02	0.17	0.02
	$\delta_1 = -1$	0.15	0.01	0.12	0.01	0.11	0.01	0.10	0.01
	$\delta_2 = 2$	0.13	0.02	0.11	0.01	0.10	0.01	0.09	0.01
	$\delta_3 = -3$	0.38	0.05	0.27	0.04	0.24	0.03	0.21	0.01

Table 4 (continued)

$n = 100$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
5	$\alpha_2 = -1$	0.26	0.03	0.26	0.03	0.27	0.03	0.25	0.03
	$\alpha_3 = -2$	0.59	0.07	0.46	0.05	0.39	0.04	0.38	0.04
	$\alpha_4 = 1$	0.37	0.04	0.30	0.03	0.30	0.03	0.29	0.03
	$\alpha_5 = 2$	0.51	0.07	0.34	0.04	0.33	0.04	0.29	0.03
	$\delta_1 = -1$	0.15	0.02	0.15	0.01	0.15	0.02	0.16	0.01
	$\delta_2 = 2$	0.21	0.02	0.18	0.02	0.17	0.02	0.14	0.01
	$\delta_3 = -3$	0.69	0.08	0.58	0.07	0.47	0.06	0.38	0.05
	$\delta_4 = 4$	0.70	0.08	0.49	0.05	0.41	0.05	0.38	0.04
	$\delta_5 = -5$	1.55	0.20	1.06	0.13	0.95	0.13	0.81	0.11

Table 5 Scenario 1: The average absolute bias and standard error obtained for each parameter over 2000 simulations when the fitted model is Model 4 ($n = 1000$)

$n = 1000$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
$m = 3$									
3	$\alpha_2 = -1$	0.19	0.02	0.14	0.01	0.14	0.01	0.13	0.01
	$\alpha_3 = 1$	0.23	0.01	0.15	0.01	0.13	0.01	0.13	0.01
	$\delta_1 = -1$	0.13	0.01	0.08	0.01	0.08	0.01	0.07	0.01
	$\delta_2 = 2$	0.16	0.01	0.12	0.01	0.10	0.01	0.09	0.01
	$\delta_3 = -3$	0.53	0.02	0.35	0.01	0.28	0.02	0.23	0.01
5	$\alpha_2 = -1$	0.40	0.02	0.31	0.01	0.24	0.01	0.23	0.01
	$\alpha_3 = -2$	0.72	0.04	0.43	0.03	0.30	0.01	0.21	0.01
	$\alpha_4 = 1$	0.55	0.02	0.37	0.01	0.29	0.01	0.26	0.01
	$\alpha_5 =2$	0.60	0.02	0.35	0.01	0.28	0.01	0.24	0.01
	$\delta_1 = -1$	0.18	0.01	0.14	0.01	0.11	0.01	0.10	0.01
	$\delta_2 = 2$	0.32	0.01	0.23	0.01	0.20	0.01	0.20	0.01
	$\delta_3 = -3$	1.70	0.01	0.78	0.04	0.50	0.03	0.33	0.01
	$\delta_4 = 4$	0.77	0.05	0.50	0.02	0.41	0.03	0.39	0.01
	$\delta_5 = -5$	1.11	0.03	0.72	0.03	0.60	0.02	0.53	0.02
$m = 5$									
3	$\alpha_2 = -1$	0.09	0.01	0.07	0.01	0.06	0.01	0.06	0.01
	$\alpha_3 = 1$	0.12	0.01	0.09	0.01	0.08	0.01	0.07	0.01
	$\delta_1 = -1$	0.06	0.01	0.05	0.01	0.05	0.01	0.04	0.01
	$\delta_2 = 2$	0.09	0.01	0.08	0.01	0.07	0.01	0.06	0.01
	$\delta_3 = -3$	0.22	0.01	0.14	0.01	0.14	0.01	0.13	0.01

Table 5 (continued)

$n = 1000$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
5	$\alpha_2 = -1$	0.16	0.01	0.13	0.01	0.12	0.01	0.12	0.01
	$\alpha_3 = -2$	0.19	0.01	0.15	0.01	0.13	0.01	0.13	0.01
	$\alpha_4 = 1$	0.23	0.01	0.21	0.01	0.17	0.01	0.16	0.01
	$\alpha_5 = 2$	0.21	0.01	0.18	0.01	0.16	0.01	0.15	0.01
	$\delta_1 = -1$	0.08	0.01	0.06	0.01	0.06	0.01	0.05	0.01
	$\delta_2 = 2$	0.17	0.01	0.14	0.01	0.13	0.01	0.12	0.01
	$\delta_3 = -3$	0.36	0.02	0.20	0.01	0.18	0.01	0.16	0.01
	$\delta_4 = 4$	0.38	0.01	0.31	0.01	0.28	0.01	0.26	0.01
	$\delta_5 = -5$	0.46	0.02	0.43	0.01	0.35	0.01	0.33	0.01
$m = 10$									
3	$\alpha_2 = -1$	0.06	0.01	0.05	0.01	0.05	0.01	0.05	0.01
	$\alpha_3 = 1$	0.07	0.01	0.06	0.01	0.06	0.01	0.05	0.01
	$\delta_1 = -1$	0.03	0.01	0.03	0.01	0.02	0.01	0.02	0.01
	$\delta_2 = 2$	0.05	0.01	0.04	0.01	0.04	0.01	0.03	0.01
	$\delta_3 = -3$	0.09	0.01	0.08	0.01	0.07	0.01	0.06	0.01
5	$\alpha_2 = -1$	0.09	0.01	0.08	0.01	0.08	0.01	0.08	0.01
	$\alpha_3 = -2$	0.09	0.01	0.09	0.01	0.08	0.01	0.08	0.01
	$\alpha_4 = 1$	0.11	0.01	0.10	0.01	0.10	0.01	0.09	0.01
	$\alpha_5 = 2$	0.13	0.01	0.10	0.01	0.09	0.01	0.09	0.01
	$\delta_1 = -1$	0.04	0.01	0.03	0.01	0.03	0.01	0.03	0.01
	$\delta_2 = 2$	0.09	0.01	0.07	0.01	0.06	0.01	0.06	0.01
	$\delta_3 = -3$	0.12	0.01	0.10	0.01	0.09	0.01	0.08	0.01
	$\delta_4 = 4$	0.19	0.01	0.16	0.01	0.16	0.01	0.13	0.01
	$\delta_5 = -5$	0.25	0.01	0.19	0.01	0.18	0.01	0.16	0.01

Table 6 Scenario 1: The average absolute bias and standard error obtained for each parameter over 2000 simulations when the fitted model is Model 3 ($n = 100$)

$n = 100$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
$m = 3$									
3	$\alpha_2 = -1$	2.23	0.19	2.08	0.15	2.01	0.14	1.86	0.09
	$\alpha_3 = 1$	2.55	0.19	2.27	0.15	2.20	0.13	1.95	0.08
5	$\alpha_2 = -1$	1.16	0.12	1.08	0.11	0.99	0.10	0.99	0.10
	$\alpha_3 = -2$	2.53	0.20	2.67	0.23	2.55	0.24	2.55	0.25
	$\alpha_4 = 1$	1.27	0.14	1.24	0.13	1.13	0.12	1.13	0.11
	$\alpha_5 = 2$	3.24	0.22	3.46	0.25	3.48	0.24	3.48	0.23
$m = 5$									
3	$\alpha_2 = -1$	1.64	0.11	1.51	0.09	1.40	0.04	1.51	0.06
	$\alpha_3 = 1$	1.72	0.09	1.57	0.01	1.58	0.03	1.60	0.03
5	$\alpha_2 = -1$	0.93	0.09	0.83	0.08	0.79	0.08	0.78	0.07
	$\alpha_3 = -2$	2.48	0.24	2.92	0.23	2.65	0.23	2.76	0.19
	$\alpha_4 = 1$	0.95	0.10	0.87	0.09	0.92	0.09	0.88	0.08
	$\alpha_5 = 2$	3.79	0.25	3.79	0.23	3.68	0.22	3.58	0.21
$m = 10$									
3	$\alpha_2 = -1$	1.17	0.04	1.16	0.02	1.19	0.02	1.23	0.02
	$\alpha_3 = 1$	1.32	0.02	1.38	0.02	1.38	0.01	1.42	0.01
5	$\alpha_2 = -1$	0.52	0.05	0.54	0.04	0.55	0.03	0.58	0.03
	$\alpha_3 = -2$	2.28	0.16	2.46	0.18	2.25	0.14	2.08	0.07
	$\alpha_4 = 1$	0.65	0.05	0.64	0.03	0.69	0.03	0.72	0.02
	$\alpha_5 = 2$	3.47	0.19	3.16	0.14	2.98	0.12	2.90	0.11

Table 7 Scenario 1: The average absolute bias and standard error obtained for each parameter over 2000 simulations when the fitted model is Model 3 (n = 1000)

$n = 1000$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	B ias	SE
$m = 3$									
3	$\alpha_2 = -1$	1.38	0.01	1.43	0.01	1.46	0.00	1.49	0.01
	$\alpha_3 = 1$	1.72	0.01	1.68	0.01	1.73	0.01	1.80	0.01
5	$\alpha_2 = -1$	0.55	0.02	0.51	0.02	0.53	0.02	0.49	0.02
	$\alpha_3 = -2$	1.55	0.02	2.01	0.03	1.99	0.04	2.07	0.03
	$\alpha_4 = 1$	0.55	0.02	0.50	0.02	0.60	0.02	0.59	0.02
	$\alpha_5 = 2$	2.20	0.02	2.50	0.04	2.83	0.04	2.89	0.04
$m = 5$									
3	$\alpha_2 = -1$	1.30	0.01	1.34	0.01	1.38	0.01	1.41	0.01
	$\alpha_3 = 1$	1.59	0.01	1.65	0.01	1.70	0.01	1.74	0.01
5	$\alpha_2 = -1$	0.44	0.02	0.54	0.01	0.60	0.01	0.70	0.01
	$\alpha_3 = -2$	2.12	0.04	2.47	0.03	2.56	0.02	2.58	0.02
	$\alpha_4 = 1$	0.48	0.02	0.57	0.01	0.64	0.01	0.73	0.01
	$\alpha_5 = 2$	2.63	0.04	2.89	0.03	3.05	0.03	3.22	0.02
$m = 10$									
3	$\alpha_2 = -1$	1.19	0.01	1.24	0.01	1.27	0.01	1.29	0.01
	$\alpha_3 = 1$	1.42	0.01	1.50	0.01	1.55	0.01	1.59	0.01
5	$\alpha_2 = -1$	0.46	0.01	0.52	0.01	0.57	0.01	0.60	0.01
	$\alpha_3 = -2$	2.21	0.02	2.19	0.01	2.23	0.01	2.26	0.01
	$\alpha_4 = 1$	0.49	0.01	0.58	0.01	0.65	0.01	0.68	0.01
	$\alpha_5 = 2$	2.63	0.02	2.70	0.01	2.79	0.01	2.83	0.01

Table 8 Scenario 2: The average absolute bias and standard error obtained for each parameter over 2000 simulations when the fitted model is Model 4 (n = 100)

$n = 100$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
$m = 3$									
3	$\alpha_2 = -1$	1.60	0.24	1.01	0.15	0.95	0.12	1.01	0.12
	$\alpha_3 = 1$	2.17	0.49	1.74	0.40	1.32	0.29	1.30	0.31
	$\delta_1 = 1$	2.23	0.14	2.11	0.05	2.06	0.05	2.06	0.05
	$\delta_2 = 2$	1.45	0.24	1.15	0.22	1.04	0.17	0.91	0.14
	$\delta_3 = 3$	10.66	2.21	8.71	1.12	7.55	0.56	7.28	0.36
5	$\alpha_2 = -1$	1.37	0.34	1.53	0.56	1.03	0.12	1.06	0.16
	$\alpha_3 = -2$	7.02	2.33	3.98	1.07	2.58	0.67	3.20	0.99
	$\alpha_4 = 1$	1.68	0.21	1.62	0.32	1.20	0.14	1.21	0.20
	$\alpha_5 = 2$	3.33	0.85	1.55	0.28	1.43	0.31	1.51	0.26
	$\delta_1 = 1$	2.15	0.06	2.12	0.06	2.14	0.05	2.17	0.07
	$\delta_2 = 2$	0.88	0.40	0.93	0.42	0.53	0.09	0.45	0.07
	$\delta_3 = 3$	22.75	10.46	8.43	0.72	8.01	0.77	14.04	6.13
	$\delta_4 = 4$	6.55	1.42	4.11	0.86	3.36	0.80	3.03	1.46
	$\delta_5 = 5$	23.72	4.44	14.65	1.25	12.32	0.45	13.75	1.03
$m = 5$									
3	$\alpha_2 = -1$	1.04	0.33	0.60	0.07	0.52	0.06	0.49	0.06
	$\alpha_3 = 1$	1.31	0.31	6.66	5.92	0.63	0.09	0.52	0.06
	$\delta_1 = 1$	2.14	0.05	2.08	0.04	2.05	0.04	2.02	0.03
	$\delta_2 = 2$	1.19	0.29	0.72	0.14	0.51	0.08	0.42	0.07
	$\delta_3 = 3$	7.84	0.70	34.25	27.85	6.39	0.09	6.21	0.08
5	$\alpha_2 = -1$	2.02	0.83	0.98	0.21	0.76	0.09	0.63	0.08
	$\alpha_3 = -2$	4.45	1.69	2.61	0.88	1.75	0.54	1.65	0.51
	$\alpha_4 = 1$	1.08	0.17	0.81	0.10	0.74	0.08	0.72	0.09
	$\alpha_5 = 2$	1.94	0.45	1.18	0.26	0.87	0.11	0.85	0.12
	$\delta_1 = 1$	2.26	0.10	2.14	0.05	2.09	0.04	2.03	0.04
	$\delta_2 = 2$	5.95	5.13	0.89	0.44	0.37	0.05	0.31	0.03
	$\delta_3 = 3$	13.04	3.37	9.65	2.09	7.15	0.65	6.98	0.65
	$\delta_4 = 4$	3.82	0.82	2.00	0.30	1.49	0.23	1.21	0.24
	$\delta_5 = 5$	15.70	1.58	12.32	0.50	11.64	0.39	11.97	0.33
$m = 10$									
3	$\alpha_2 = -1$	0.55	0.27	0.25	0.03	0.21	0.02	0.20	0.02
	$\alpha_3 = 1$	0.43	0.05	0.31	0.04	0.25	0.03	0.26	0.03
	$\delta_1 = 1$	2.09	0.02	2.07	0.02	2.06	0.02	2.05	0.02
	$\delta_2 = 2$	0.82	0.31	0.31	0.05	0.25	0.03	0.20	0.02
	$\delta_3 = 3$	6.29	0.07	6.09	0.05	6.02	0.04	6.02	0.04

Table 8 (continued)

$n = 100$

R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
5	$\alpha_2 = -1$	0.46	0.05	0.40	0.05	0.39	0.05	0.31	0.04
	$\alpha_3 = -2$	0.93	0.14	1.45	0.80	0.50	0.06	0.54	0.11
	$\alpha_4 = 1$	0.50	0.05	0.48	0.05	0.49	0.05	0.41	0.04
	$\alpha_5 = 2$	0.83	0.12	0.60	0.09	0.48	0.07	0.47	0.06
	$\delta_1 = 1$	2.12	0.03	2.08	0.02	2.09	0.02	2.06	0.02
	$\delta_2 = 2$	0.36	0.04	0.26	0.03	0.25	0.03	0.25	0.03
	$\delta_3 = 3$	6.75	0.32	7.47	1.25	6.07	0.09	6.10	0.10
	$\delta_4 = 4$	1.09	0.15	0.85	0.11	0.69	0.09	0.60	0.07
	$\delta_5 = 5$	7.26	0.22	6.86	0.15	6.59	0.10	5.76	0.14

Table 9 Scenario 2: The average absolute bias and standard error obtained for each parameter over 2000 simulations when the fitted model is Model 4 (n=1000)

$n = 1000$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
$m = 3$									
3	$\alpha_2 = -1$	0.35	0.02	0.28	0.01	0.20	0.01	0.19	0.01
	$\alpha_3 = 1$	0.40	0.02	0.30	0.01	0.27	0.01	0.25	0.01
	$\delta_1 = 1$	1.99	0.01	2.00	0.01	2.00	0.01	2.00	0.01
	$\delta_2 = 2$	0.46	0.02	0.50	0.02	0.42	0.01	0.31	0.01
	$\delta_3 = 3$	0.53	0.04	0.51	0.01	0.59	0.03	0.52	0.01
5	$\alpha_2 = -1$	0.43	0.02	0.37	0.01	0.32	0.02	0.32	0.01
	$\alpha_3 = -2$	0.59	0.04	0.44	0.02	0.50	0.05	0.50	0.02
	$\alpha_4 = 1$	0.57	0.02	0.49	0.01	0.45	0.02	0.44	0.01
	$\alpha_5 = 2$	0.94	0.04	0.63	0.03	0.59	0.03	0.45	0.02
	$\delta_1 = 1$	0.92	0.01	0.94	0.01	0.92	0.01	0.96	0.01
	$\delta_2 = 2$	0.36	0.01	0.35	0.01	0.37	0.01	0.33	0.01
	$\delta_3 = 3$	0.37	0.05	0.31	0.03	0.33	0.04	0.46	0.03
	$\delta_4 = 4$	0.73	0.07	0.54	0.03	0.77	0.05	0.48	0.02
	$\delta_5 = 5$	0.73	0.10	0.92	0.05	0.62	0.08	0.61	0.04
$m = 5$									
3	$\alpha_2 = -1$	0.16	0.01	0.13	0.01	0.14	0.01	0.12	0.01
	$\alpha_3 = 1$	0.18	0.01	0.16	0.01	0.14	0.01	0.14	0.01
	$\delta_1 = 1$	0.98	0.01	0.98	0.01	0.99	0.01	0.99	0.01
	$\delta_2 = 2$	0.26	0.01	0.20	0.01	0.17	0.01	0.15	0.01
	$\delta_3 = 3$	0.36	0.01	0.34	0.01	0.34	0.03	0.33	0.01
5	$\alpha_2 = -1$	0.30	0.01	0.28	0.01	0.24	0.01	0.25	0.01
	$\alpha_3 = -2$	0.93	0.13	0.40	0.02	0.38	0.02	0.41	0.02
	$\alpha_4 = 1$	0.43	0.02	0.45	0.02	0.42	0.02	0.39	0.01
	$\alpha_5 = 2$	0.86	0.09	0.58	0.03	0.53	0.03	0.47	0.02
	$\delta_1 = 1$	0.97	0.01	0.94	0.01	0.94	0.01	0.95	0.01
	$\delta_2 = 2$	0.37	0.01	0.26	0.01	0.25	0.01	0.25	0.01
	$\delta_3 = 3$	0.63	0.12	0.32	0.02	0.32	0.02	0.43	0.03
	$\delta_4 = 4$	0.82	0.04	0.59	0.02	0.48	0.02	0.49	0.02
	$\delta_5 = 5$	0.38	0.16	0.86	0.07	0.84	0.07	0.68	0.04
$m = 10$									
3	$\alpha_2 = -1$	0.07	0.01	0.06	0.01	0.05	0.01	0.05	0.01
	$\alpha_3 = 1$	0.08	0.01	0.07	0.01	0.06	0.01	0.06	0.01
	$\delta_1 = 1$	0.97	0.01	0.98	0.01	0.98	0.01	0.98	0.01
	$\delta_2 = 2$	0.09	0.01	0.08	0.01	0.06	0.01	0.06	0.01
	$\delta_3 = 3$	0.31	0.01	0.31	0.01	0.34	0.01	0.33	0.01

Table 9 (continued)

$n = 1000$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
5	$\alpha_2 = -1$	0.22	0.01	0.18	0.01	0.15	0.01	0.14	0.01
	$\alpha_3 = -2$	0.38	0.01	0.23	0.01	0.16	0.01	0.14	0.01
	$\alpha_4 = 1$	0.34	0.01	0.34	0.01	0.30	0.01	0.27	0.01
	$\alpha_5 = 2$	0.50	0.03	0.26	0.01	0.21	0.01	0.17	0.01
	$\delta_1 = 1$	0.96	0.01	0.97	0.01	0.97	0.01	0.98	0.01
	$\delta_2 = 2$	0.24	0.01	0.18	0.01	0.14	0.01	0.27	0.01
	$\delta_3 = 3$	0.38	0.02	0.23	0.02	0.31	0.01	0.14	0.01
	$\delta_4 = 4$	0.13	0.02	0.33	0.01	0.34	0.01	0.23	0.01
	$\delta_5 = 5$	0.11	0.14	0.32	0.03	0.19	0.01	0.13	0.01

Table 10 Scenario 2: The average absolute bias and standard error obtained for each parameter over 2000 simulations when the fitted model is Model 3 (n = 100)

$n = 100$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE
$m = 3$									
3	$\alpha_2 = -1$	2.62	0.19	2.10	0.15	1.93	0.10	1.91	0.10
	$\alpha_3 = 1$	2.74	0.18	2.21	0.14	2.20	0.12	2.47	0.14
5	$\alpha_2 = -1$	1.33	0.14	1.17	0.12	1.16	0.12	1.07	0.11
	$\alpha_3 = -2$	2.31	0.18	2.72	0.24	2.74	0.24	3.25	0.24
	$\alpha_4 = 1$	1.29	0.14	1.25	0.13	1.09	0.11	1.07	0.11
	$\alpha_5 = 2$	3.41	0.21	3.31	0.24	2.86	0.22	3.20	0.22
$m = 5$									
3	$\alpha_2 = -1$	1.55	0.05	1.42	0.02	1.48	0.03	1.55	0.05
	$\alpha_3 = 1$	2.20	0.15	1.92	0.05	1.93	0.07	1.97	0.06
5	$\alpha_2 = -1$	0.97	0.10	0.90	0.09	0.91	0.09	0.87	0.08
	$\alpha_3 = -2$	2.88	0.25	2.08	0.19	2.62	0.24	2.52	0.19
	$\alpha_4 = 1$	0.93	0.10	0.87	0.09	0.87	0.08	0.86	0.07
	$\alpha_5 = 2$	3.43	0.25	3.48	0.23	3.13	0.19	3.46	0.21
$m = 10$									
3	$\alpha_2 = -1$	1.26	0.02	1.29	0.01	1.32	0.01	1.33	0.02
	$\alpha_3 = 1$	1.59	0.02	1.61	0.02	1.62	0.02	1.66	0.02
5	$\alpha_2 = -1$	0.65	0.05	0.69	0.05	0.72	0.03	0.74	0.03
	$\alpha_3 = -2$	2.33	0.17	2.16	0.12	2.14	0.11	2.12	0.11
	$\alpha_4 = 1$	0.64	0.06	0.65	0.04	0.65	0.04	0.77	0.03
	$\alpha_5 = 2$	2.81	0.17	2.63	0.12	2.56	0.10	2.62	0.11

Table 11 Scenario 2: The average absolute bias and standard error obtained for each parameter over 2000 simulations when the fitted model is Model 3 ($n = 1000$)

$n = 1000$									
R	True parameter	$q = 3$		$q = 4$		$q = 5$		$q = 6$	
		Bias	SE	Bias	SE	Bias	SE	bias	SE
$m = 3$									
3	$\alpha_2 = -1$	1.50	0.01	1.47	0.01	1.53	0.01	1.56	0.01
	$\alpha_3 = 1$	1.70	0.02	1.68	0.01	1.76	0.02	1.83	0.01
5	$\alpha_2 = -1$	0.61	0.02	0.55	0.02	0.57	0.02	0.59	0.02
	$\alpha_3 = -2$	1.61	0.02	1.89	0.03	2.17	0.03	2.39	0.03
	$\alpha_4 = 1$	0.49	0.02	0.56	0.02	0.62	0.02	0.61	0.02
	$\alpha_5 = 2$	2.03	0.03	2.71	0.04	2.72	0.04	2.77	0.03
$m = 5$									
3	$\alpha_2 = -1$	1.36	0.01	1.41	0.01	1.45	0.01	1.49	0.01
	$\alpha_3 = 1$	1.53	0.01	1.58	0.01	1.63	0.01	1.69	0.01
5	$\alpha_2 = -1$	0.52	0.01	0.61	0.01	0.68	0.01	0.72	0.01
	$\alpha_3 = -2$	2.21	0.04	2.46	0.02	2.48	0.05	2.40	0.02
	$\alpha_4 = 1$	0.46	0.02	0.57	0.02	0.66	0.02	0.76	0.01
	$\alpha_5 = 2$	2.39	0.04	2.69	0.03	3.11	0.03	3.12	0.02
$m = 10$									
3	$\alpha_2 = -1$	1.26	0.01	1.30	0.01	1.34	0.01	1.37	0.01
	$\alpha_3 = 1$	1.37	0.01	1.44	0.01	1.52	0.01	1.56	0.01
5	$\alpha_2 = -1$	0.50	0.01	0.57	0.01	0.64	0.01	0.69	0.01
	$\alpha_3 = -2$	2.08	0.01	2.12	0.01	2.20	0.01	2.22	0.01
	$\alpha_4 = 1$	0.45	0.01	0.54	0.01	0.59	0.01	0.65	0.01
	$\alpha_5 = 2$	2.54	0.01	2.61	0.01	2.66	0.01	2.73	0.01

B Application: comparison of clustering structures

See Tables 12 and 13.

Table 12 Arthritis data set: Comparison of clustering structure agreement between models without covariates (3) (left) versus with covariates (4) (right)

R	Clustering comparison	ARI	1-NVI	1-NID
2	No covariate versus age	0.86	0.69	0.81
	No covariate versus treatment	0.74	0.51	0.68
	No covariate versus gender	0.87	0.70	0.82
	No covariate versus age, treatment	0.69	0.45	0.61
	No covariate versus age, gender	0.87	0.70	0.82
	No covariate versus treatment, gender	0.75	0.55	0.70
	No covariate versus age, treatment, gender	0.70	0.46	0.62
3	No covariate versus age	0.43	0.37	0.52
	No covariate versus treatment	0.14	0.12	0.21
	No covariate versus gender	0.33	0.32	0.47
	No covariate versus age, treatment	0.58	0.25	0.38
	No covariate versus age, gender	0.70	0.61	0.74
	No covariate versus treatment,gender	0.27	0.23	0.36
	No covariate versus age, treatment, gender	0.14	0.17	0.27
4	No covariate versus age	0.44	0.34	0.48
	No covariate versus treatment	0.85	0.68	0.80
	No covariate versus gender	1.00	1.00	1.00
	No covariate versus age, treatment	**0.66**	**0.47**	**0.64**
	No covariate versus age, gender	0.76	0.57	0.71
	No covariate versus treatment, gender	0.84	0.67	0.80
	No covariate versus age, treatment, gender	0.58	0.39	0.55

Table 13 The row clustering structure of models without covariates (3) and with covariates (4), when $R = 4$, x_1 : age and x_2 : treatments. The differences between models in clustering structure are highlighted in boldface

Cluster	$\mu_k - \alpha_r$	$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$
1	**1** 14 24 54 61 **62** 67 100 **119** **124** **131** 134 137 141 **223** **239** **243** **266**	14 24 54 61 67 100 134 137 141
2	2 3 4 5 11 12 17 23 26 27 29 30 32 33 35 37 38 42 43 46 47 50 52 58 59 65 68 69 70 71 **79** 81 82 83 85 90 92 94 101 102 104 106 107 109 113 116 117 120 121 126 133 140 143 150 157 161 163 168 170 172 173 180 183 185 186 187 191 193 199 200 201 205 206 207 208 211 212 213 214 222 225 226 227 229 234 237 **238** 248 255 257 258 263 269 272 273 281	**1** 2 3 4 5 11 12 17 23 26 27 29 30 32 33 35 37 38 42 43 46 47 50 52 58 59 **62** 65 68 69 70 71 81 82 83 85 90 92 94 101 102 104 106 107 109 113 116 117 **119** 120 121 **124** 126 **131** 133 140 143 150 157 161 163 168 170 172 173 180 183 185 186 187 191 193 199 200 201 205 206 207 208 211 212 213 214 222 **223** 225 226 227 229 234 237 **239** **243** 248 255 257 258 263 **266** 269 272 273 281
3	6 7 8 10 13 15 16 19 20 21 25 28 31 34 36 39 40 41 44 45 48 49 51 53 56 57 **63** 64 66 73 74 75 76 78 80 84 86 87 88 89 91 93 96 97 98 99 103 105 108 110 111 112 114 115 118 122 123 **125** 127 128 129 130 135 136 138 139 142 144 145 146 147 148 149 151 152 **153** 154 155 156 158 160 162 164 165 166 167 169 174 176 177 178 179 181 184 189 190 **192** 194 195 196 197 198 204 209 210 **215** **217** 218 **219** 220 221 224 228 230 231 232 233 235 236 240 242 244 **245** 247 251 252 253 254 256 259 260 261 262 264 **267** 270 271 274 275 276 277 278 279 280 282 283 284 **285** 289	6 7 8 10 13 15 16 19 20 21 25 28 31 34 36 39 40 41 44 45 48 49 51 53 56 57 64 66 73 74 75 76 78 **79** 80 84 86 87 88 89 91 93 96 97 98 99 103 105 108 110 111 112 114 115 118 122 123 127 128 129 130 135 136 138 139 142 144 145 146 147 148 149 151 152 154 155 156 158 160 162 164 165 166 167 169 174 176 177 178 179 181 184 189 190 194 195 196 197 198 204 209 210 218 220 221 224 228 230 231 232 233 235 236 **238** 240 242 244 247 251 252 253 254 256 259 260 261 262 264 270 271 274 275 276 277 278 279 280 282 283 284 289
4	9 18 22 55 60 72 77 95 132 159 171 175 182 188 202 203 216 241 246 249 250 265 268 286 287 288	9 18 22 55 60 **63** 72 77 95 **125** 132 **153** 159 171 175 182 188 **192** 202 203 **215** 216 **217** **219** 241 **245** 246 249 250 265 **267** 268 **285** 286 287 288

C Application: comparison among models with *R*=4 clusters

See Table 14.

Table 14 Average age for each cluster (R = 4) and for each level of treatment (placebo or drug), for the model with covariates age and treatment ($\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$, x_1 = age & x_2 = treatment) and without covariates ($\mu_k - \alpha_r$)

Treatment	R1		R2		R3		R4	
	Placebo	Drug	Placebo	Drug	Placebo	Drug	Placebo	Drug
$\mu_k - \alpha_r$	45.5	35.0	54.5	50.2	53.5	53.6	57.4	52.5
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$	46.5	41.0	50.8	49.2	50.5	51.1	53.4	50.1

D Application: comparison with Partitioning Around Medoids (PAM)

See Table 15.

Table 15 Average age for each cluster (R = 4) and for each level of treatment (placebo or drug), for the model with covariates age and treatment ($\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$, x_1 = age & x_2 = treatment) and the Partitioning Around Medoids (PAM)

Treatment	R1		R2		R3		R4	
	Placebo	Drug	Placebo	Drug	Placebo	Drug	Placebo	Drug
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$ $\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$	46.5	41.0	50.8	49.2	50.5	51.1	53.4	50.1
PAM	45.8	46.0	51.4	49.9	51.1	51.7	51.3	50.1

E Application: robustness analysis

To evaluate the robustness of our proposal, we artificially and randomly introduced 3% outliers in the numerical variable “age" and assessed the model fitting and its performance in capturing the underlying clustering structure (Table 16).

Table 16 Results of row clustering models fitted to the arthritis data set with outliers The best model in each group of models (one, two, or three covariates), based on AIC and BIC, is shown in bold

Model		R	Number of parameter	Log-like	AIC	BIC
$\mu_k - (\alpha_r + x_i\delta_r)$	$x = \text{age}$	2	8	−1128.95	2273.89	2312.01
		3	11	−1095.13	2212.26	**2264.68**
		4	14	−1095.90	2219.81	2286.52
		5	17	−1079.59	**2193.18**	2274.19
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$	$x_1 = \text{age}$	2	10	−1105.94	2231.87	2279.52
	$x_2 = \text{treatment}$	3	14	−1078.26	2184.51	**2251.22**
		4	18	−1080.34	2196.67	2282.44
		5	22	−1064.40	**2172.79**	2277.62
	$x_1 = \text{age}$	2	10	−1138.75	2297.49	2345.14
	$x_2 = \text{gender}$	3	14	−1095.96	2219.92	**2286.63**
		4	18	−1088.36	2212.72	2298.49
		5	22	−1083.55	**2211.10**	2315.93
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r} + x_{i3}\delta_{3r})$	$x_1 = \text{age}$	2	12	−1108.36	2240.72	2297.90
	$x_2 = \text{ treatment}$	3	17	−1080.82	2195.64	**2276.64**
	$x_3 = \text{gender}$	4	22	−1097.13	2238.26	2343.09
		5	27	−1063.01	**2180.02**	2308.68

Acknowledgements This work has been supported by the Ministerio de Ciencia e Innovación (Spain) [PID2019-104830RB-I00/ DOI (AEI): 10.13039/501100011033], and by Grant 2021 SGR 01421 (GRBIO) administrated by the Departament de Recerca i Universitats de la Generalitat de Catalunya (Spain). Daniel Fernández is member of the Centro de Investigación Biomédica en Red de Salud Mental, Instituto de Salud Carlos III (CIBERSAM). Daniel Fernández is a Serra Húnter Fellow.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest The authors declare that they have no conflict of interest. Our research do not use or survey participants. All ethical conduct rules were followed. The analysed data used during the current study is obtained from the **R** package *multgee* (Touloumis 2015).

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

Agresti A (2014) Analysis of ordinal categorical data, 3rd edn. John Wiley and Sons Inc (Wiley Series in Probability and Statistics)

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Proceedings of the 2nd international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281

Anderson JA (1984) Regression and ordered categorical variable. J R Stat Soc 46:1–30

Andrews JL, McNicholas PD (2014) Variable selection for clustering and classification. J Classif 31(2):136–153

Biernacki C, Jacques J (2016) Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Stat Comput 26:929–943

Böhning D, Seidel W, Alfó M, Garel B, Patilea V, Walther G (2007) Advances in mixture models. Comput Stat Data Anal 51(11):5205–5210

Bouveyron C, Brunet-Saumard C (2014) Model-based clustering of high-dimensional data: a review. Comput Stat Data Anal 71:52–78

Brant R (1990) Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics 1171–1178

Clogg CC (1988) Latent class models for measuring. Latent trait and latent class models, pp 173–205

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1(1–4):131–156

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38

Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5(Aug):845–889

Everitt B, Landau S, Leese M, Stahl D (2011) Clust Anal. John Wiley and Sons, New York

Fernández D, Arnold R, Pledger S (2016) Mixture-based clustering for the ordered stereotype model. Comput Stat Data Anal 93:46–75

Fernández D, Arnold R, Pledger S, Liu I, Costilla R (2019) Finite mixture biclustering of discrete type multivariate data. Adv Data Anal Classif 13:117–143

Formann AK (1992) Linear logistic latent class analysis for polytomous data. J Am Stat Assoc 87(418):476–486

Fowlkes EB, Gnanadesikan R, Kettenring JR (1988) Variable selection in clustering. J Classif 5(2):205–228

Giordan M, Diana G (2011) A clustering method for categorical ordinal data. Commun Stat Theory Methods 40(7):1315–1334

Govaert G, Nadif M (2010) Latent block model for contingency table. Commun Stat Theory Methods 39(3):416–425

Gudicha DW, Vermunt JK (2013) Mixture model clustering with covariates using adjusted three-step approaches. In: Algorithms from and for nature and life. Springer, pp 87–94

Hartigan JA, Wong MA (1979) A k-means clustering algorithm. Appl Stat 28:100–108

Hennig C (2015) What are the true clusters? Patt Recogn Lett 64:53–62. https://doi.org/10.1016/j.patrec.2015.04.009

Hoff PD (2005) Subset clustering of binary sequences, with an application to genomic abnormality data. Biometrics 61:1027–1036

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Ingrassia S, Minotti SC, Vittadini G (2012) Local statistical modeling via a cluster-weighted approach with elliptical distributions. J Classif 29(3):363–401

Ingrassia S, Punzo A, Vittadini G, Minotti SC (2015) Erratum to: the generalized linear mixed cluster-weighted model. J Classif 32(2):327–355

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3(1):79–87

John GH, Kohavi R, Pfleger K (1994) Irrelevant features and the subset selection problem. In: Machine learning proceedings 1994. Elsevier, pp 121–129

Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32:241–254

Kraskov A, Stögbauer H, Andrzejak R, Grassberger P (2005) Hierarchical clustering using mutual information. EPL 70:278–284

Lamont AE, Vermunt JK, Van Horn ML (2016) Regression mixture models: does modeling the covariance between independent variables and latent classes improve the results? Multivar Behav Res 51(1):35–52

Lipsitz SR, Fitzmaurice GM, Molenberghs G (1996) Goodness-of-fit tests for ordinal response regression models. J R Stat Soc Ser C 45(2):175–190

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Proc Berkeley Symp Math Stat Probab 1:281–297

Matechou E, Liu I, Fernández D, Farias M, Gjelsvik B (2016) Biclustering models for two-mode ordinal data. Psychometrika 81(3):611–624

McCullagh P (1980) Regression models for ordinal data. J R Stat Soc 42:109–142

McLachlan G, Basford K (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York

McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New york

Meila M (2005) Comparing clusterings: an axiomatic view. ACM Press, pp 577–584

Meila M (2007) Comparing clusterings: an information based distance. J Multivar Anal 98:873–895

Melnykov V, Maitra R (2010) Finite mixture models and model-based clustering. Stat Surv 4:80–116

Müller P, Quintana F, Rosner GL (2011) A product partition model with regression on covariates. J Comput Graph Stat 20:1:260–278. https://doi.org/10.1198/jcgs.2011.09066

Murphy K, Murphy TB (2020) Gaussian parsimonious clustering models with covariates and a noise component. Adv Data Anal Class 14:293–325. https://doi.org/10.1007/s11634-019-00373-8

Peel D, McLachlan G (2000) Finite mixture models. John Wiley and Sons Inc (Wiley Series in Probability and Statistics)

Pledger S, Arnold R (2014) Multivariate methods using mixtures: correspondence analysis, scaling and pattern-detection. Comput Stat Data Anal 71:241–261

Raftery AE, Dean N (2006) Variable selection for model-based clustering. J Am Stat Assoc 101(473):168–178

Ranalli M, Rocci R (2016) Mixture models for ordinal data: a pairwise likelihood approach. Stat Comput 26:529–547

Ranalli M, Rocci R (2017) Mixture models for mixed-type data through a composite likelihood approach. Comput Stat Data Anal 110:87–102

Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat 66:846–850

Rocci R, Vichi M (2008) Two-mode multi-partitioning. Comput Stat Data Anal 52:1984–2003

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Touloumis A (2015) R package multgee: a generalized estimating equations solver for multinomial responses. J Stat Softw 64(8):1–14

Vichi M (2001) Double k-means clustering for simultaneous classification of objects and variables. In: Borra S, Rocci R, Vichi M, Schader M (eds) Adv Classif Data Anal. Springer, Berlin Heidelberg, pp 43–52

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244

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